ABSTRACT

Title of Dissertation: EXAMINING DIFFERENTIAL ITEM FUNCTIONING FROM A LATENT CLASS PERSPECTIVE

Karen Samuelsen, Doctor of Philosophy, 2005

Dissertation directed by: C. Mitchell Dayton
Department of Measurement, Statistics and Evaluation

Current approaches for studying differential item functioning (DIF) using manifest groups are problematic since these groups are treated as homogeneous in nature. Additionally, manifest variables such as sex and ethnicity are proxies for more fundamental differences – educational advantage/disadvantage attributes. A simulation study was conducted to highlight issues arising from the use of standard DIF detection procedures. Results of this study showed that as the amount of overlap between manifest groups and latent classes decreased, so did the power to correctly identify items with DIF. Furthermore, the true magnitude of the DIF was obscured making it increasingly more difficult to eliminate items on that basis.

After some problems with manifest group approaches for DIF had been identified, a recovery study was conducted using the WINBUGS software in the analysis of the mixed Rasch model for detecting DIF. In this study the mixed Rasch model also showed a lack of power to detect items with DIF when the sample size was small. However, this

approach was able to identify the proportion of and ability distribution for each manifest group within latent classes, thereby providing a mechanism for judging the appropriateness of using manifest variables as proxies for latent ones. Finally, a series of protocols was developed for examining DIF using a latent class approach, and these were used to examine differential item functioning on a test of language proficiency for English language learners. Results showed that 74% of Hispanic and 83% of Asian examinees were in one latent class, meaning any DIF found by comparing manifest groups would be an artifact of a relatively small number of examinees. Examination of the output from the latent class analysis provided potentially important insights into the causes of DIF, however covariates were not predictive of latent class membership.

EXAMINING DIFFERENTIAL ITEM FUNCTIONING FROM A LATENT CLASS PERSPECTIVE

by

Karen Samuelsen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor C. Mitchell Dayton, Chair
Professor Paul Hanges
Professor Amy Hendrickson
Professor George Macready
Professor Robert Mislevy

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Differential item functioning (DIF) occurs "when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest" (Clauser & Mazor, 1998, pg. 31). The presence of DIF means that scores from different groups are not comparable – a fact that compromises the inferences made regarding examinees. Differential item functioning also signals multidimensionality due to the presence of nuisance dimensions (Ackerman, 1992). The presence of DIF calls into question the inferences drawn through the unidimensional models most often used in measurement. Finally, at a more fundamental level, the presence of DIF raises issues regarding fairness and equity in testing.

Because DIF does have serious consequences, it has generated extensive research. This research has focused largely on psychometric concerns such as the appropriateness of the matching criterion, the question of whether the item of interest should be included in the matching criterion, and which procedures work best under a variety of conditions (Clauser & Mazor, 1998). While these issues are important, there are two major problems with the current procedures for identifying DIF that have been rarely studied. The first issue relates to the use of manifest grouping variables, such as sex, race and ethnicity. The second related issue deals with the lack of information gained regarding the cause of the DIF. The purpose of this research is to illuminate these issues in some detail and offer solutions to those currently studying DIF and those concerned about identifying items on operational assessments that function differentially.

It has been stated that "traditional DIF analyses are based on the *de facto* assumption that individuals within a manifest group are more similar to one another than they are to members of the other manifest group" (DeAyala, Kim, Stapleton & Dayton, 2002, pg. 247). In reality, although genders, racial groups, and ethnic groups are easily identified, they often do not represent homogeneous populations. A widely used example showing this variability within groups is the Hispanic population in the United States. According to the US Census Bureau (1993), "Persons of Hispanic origin, in particular, were those who indicated that their origin was Mexican, Puerto Rican, Cuban, Central or South American, or some other Hispanic origin. It should be noted that persons of Hispanic origin might be of any race." Given this diversity in place of origin and race, it seems obvious that a classification of Hispanic will yield a heterogeneous group. The same can be said for classifications based on other ethnicities, race or sex. Based on this lack of homogeneity, when items demonstrate DIF with regard to a manifest group, a portion of the subjects need not be expected to respond like other members of their group. In a study by Cohen and Bolt (2002), items were examined that demonstrated DIF across gender. When these items were re-examined using a latent class approach, over half of the females in the study were assigned to the opposite group that their gender would indicate; the same was true for nearly 40% of the males. This illustrates just how tenuous the relationship can be between the manifest groups used to examine DIF and the latent groups whom the items truly advantage or disadvantage.

In addition to the lack of homogeneity in these groups, there is also the possibility that the groups being examined are not really the manifest groups affected. Hu and Dorans (1989) found, as would be expected, that the removal of an item favoring females

resulted in slightly lower scores for females and slightly higher scores for males. What surprised the researchers was that the scores of both Hispanics and Asian-Americans were raised more than the scores of males, meaning that females in those groups actually received an advantage by the removal of the item. These results demonstrated a flaw in DIF analyses that concentrate on the marginals and point to the need to consider interactions in DIF analyses. Dorans and Holland (1993, pg. 64) introduced the idea of "Melting-Pot DIF" analyses to solve this problem by comparing item function for each gender/ethnic group to the population of all other examinees (the melting pot). Their solution seems to have found little support, possibly because it would increase the number of required analyses, and it would be more difficult to find DIF in larger groups (e.g. white males or white females) using this strategy.

Another reason not to use manifest groups for DIF is they are not directly related to the issues of learning educators care about. Researchers have long argued that manifest groups defined by characteristics like sex and ethnicity are really proxies for something else. Dorans and Holland (1993, pg. 65) wrote:

> "It could be argued, however, that these intact ethnic groups are merely surrogates for an educational disadvantage attribute that should be used in focal group definition. In fact, within any of these groups, there is probably a considerable degree of variability with respect to educational advantage or disadvantage. Perhaps we should be focusing our group definition efforts toward defining and measuring educational advantage or disadvantage directly."

An important question to ask is if these manifest grouping variables are, in fact, surrogates for other attributes, what happens when they are used in lieu of those attributes? One possibility is that we miss items that are functioning differentially based on this latent attribute but not based on the manifest grouping variable. This has ramifications with regard to validity arguments and the ability to understand the causes of

the differential function. Another possibility is that we incorrectly assume an item or items exhibiting DIF disadvantage **all** members of a manifest group. DeAyala, et al. (2002) found that black examinees in one latent class were affected by three test items though black examinees in the other class were not. As those authors note, in order to state that a manifest group is disadvantaged we would need to find items that "exhibit DIF regardless of the latent class" (pg. 273).

A third possibility for items that do exhibit DIF is that the true magnitude of the differential functioning may be obscured due to the lack of overlap between the manifest groups and the latent classes. This has consequences for testing companies because they treat this "observed DIF" as though it were the truth and, in the case of Educational Testing Service (ETS), use the magnitude of the DIF in classifying items into three categories. Those classifications are then used in the selection of items for operational tests (Zieky, 1993). In addition, the classification of items regarding differential function is always a precursor to discovering which items are biased and removing them.

Despite the above issues regarding the use of manifest grouping variables, they are still commonly used in DIF analyses. According to DeAyala, et al. (2002, pg. 274) that is because "The selection of manifest grouping variables is based on political not psychometric considerations". Assuming, for the sake of argument, that is the reason, a case can be made that using a latent class approach can both meet psychometric demands and satisfy political realities. From a psychometric point of view, a latent class approach, in which group differences are maximized, allows researchers to accurately capture the presence and magnitude of the differential functioning. At the same time it is possible to map manifest groups onto latent classes to satisfy those who require that connection.

After it has been established that some items do function differentially, there is a need to determine the source of that DIF. Investigations into the cause of DIF have mainly relied on statistical analyses followed by reviews of experts examining the content for obvious causes of DIF or searching for patterns that might suggest the identity of a nuisance dimension underlying it. Many agree these methodologies have had limited success in clearly defining why items function differentially (O'Neill & McPeak, 1993; Camilli & Shepard, 1994; Roussos & Stout, 1996b; Gierl, Bisanz, Bisanz, Boughton & Khaliq, 2001). Perhaps the most succinct commentary on how inadequate our current methodologies are came in the 1999 Standards for Educational and Psychological Testing.

> "Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee group to another, it has been difficult to specify the reasons for the differential performance or to identify the common deficiency among the identified items." (1999, pg. 78)

Some steps have been made towards identifying the causes of DIF using approaches that examine aspects of the items in question. Green, Crone and Folk (1989) developed an observed-score method for assessing differential distractor functioning (DDF). Thissen, Steinberg and Wainer (1993) followed up with an IRT-based methodology they called differential alternative functioning (DAF). Though both of these approaches help to pinpoint where in the item different examinees are making mistakes, they still cannot fully explain why these mistakes happen. Some researchers have also used differential speededness (Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990) and differential omission (Rivera & Schmitt, 1988; Schmitt & Dorans, 1990) to

help describe the DIF, but these are also of little help in explaining why it is occurring. More recently Roussos and Stout (1996b) developed a confirmatory, two-stage approach to help explain the underlying causes of the DIF based on Ackerman's (1992) notion of items with DIF eliciting one or more secondary dimensions beyond the primary one of interest. This procedure begins with a substantive analysis to identify items or groups of items having specified characteristics and ends with a statistical analysis (or analyses) to test whether the data reveals the dimensions indicated by the substantive analysis.

The underlying theme of this research is that it is more appropriate to use a latent conceptualization of DIF. Besides being consistent with a view of DIF stemming from multidimensionality, this conceptualization helps clarify which items function differently and can provide insights into the cause of the DIF. Kelderman and Macready (1990) agree with the idea that a latent class approach to DIF could be productive in that regard. "The use of latent grouping variables……allows for the assessment of DIF without tying that DIF to any specific variable or set of variables. Thus, it may be possible following the investigation of DIF to make a more definitive statement regarding its presence" (pg. 309).

Using a latent class approach would help explain why the differential function is occurring in two ways. First, since all items functioning differentially would be identified, along with a truer indication of the magnitude of the DIF, there is more information available to the researcher. Current strategies may only identify a subset of the items functioning differentially and underestimate the magnitude of the DIF, making it more difficult to isolate the cause of the differential functioning. Second, a latent strategy would allow researchers to incorporate covariates as predictors of group

membership. This would provide more information regarding the underlying cause of differential item function without running excessive numbers of DIF analyses. There are three specific types of predictors that may prove interesting as covariates – they are non-traditional manifest groups, interactions between traditional and non-traditional manifest groups, and continuous covariates. Examples of what can be called non-traditional manifest groups could be dichotomies such as non-native speakers versus native speakers, urban vs. non-urban students, or students taught using phonics versus those taught using whole language, to name a few. As the research by Hu and Dorans (1989) showed, interactions between manifest groups may actually be informative in terms of DIF. Likewise, interactions between sex or race, and the non-traditional groups highlighted above, may be interesting. For example, it could be that non-native speakers of Spanish are advantaged on items that include words that are somewhat uncommon in English, but which have a root that is very common in Spanish. The final category of predictor, continuous covariates, is particularly interesting since current approaches for studying DIF generally do not accommodate this type of data. Examples of continuous covariates in the field of education could be the number of math classes a student has taken or the number of years an English language learner has been in the United States.

Two separate but necessary lines of research involving latent class analyses of DIF were examined in this study. The first involved providing evidence that examining items for differential functioning using manifest groups is problematic when members of the manifest groups are not homogeneous in terms of the secondary dimension causing the DIF. This line of research built upon previous work of other researchers (Cohen & Bolt, 2002; DeAyala, et al., 2002) and further elucidates the loss of power, inflated Type

I error rate, and underestimation of the magnitude of the DIF when the manifest groups and latent classes do not completely overlap.

The second line of research involved establishing protocols for and then applying a latent class DIF approach to test data from an operational assessment of English language proficiency. Given the exploratory nature of this study, it was important to anchor this research back to the existing methodologies using manifest grouping variables to provide a means of comparison between these opposing strategies. To accomplish this, the percentage of examinees from four manifest groups (male, female, white, Hispanic) within each latent class were determined. In addition, sex and ethnicity were entered as predictors of latent class membership and significance was examined. Finally, other covariates were employed to describe the latent classes and in doing so explain why the differential functioning is occurring. The rationales for this second line of research are threefold. First, the point can be made that the manifest groups commonly used do not map well onto the latent groups tapping into educational advantage and disadvantage. Second, it can be shown that this methodology can be helpful in determining why DIF is occurring. Finally, this demonstration will show that it is possible to implement and interpret the results from a latent class approach to DIF detection.

Chapter 2: Theoretical Background

This research applied a standard DIF detection technique, the Mantel-Haenszel procedure, and a latent class approach using a mixed Rasch model. This chapter provides the rationales for the choice of these models as well as a discussion of the theoretical underpinnings of each approach. In addition, the technique used for parameter estimation, Markov chain Monte Carlo using the WINBUGS software (Spiegelhalter, Thomas & Best, 2000), is described.

Mantel-Haenszel Procedure

To demonstrate the shortcomings of the current strategy of using manifest groups in DIF studies, it is necessary to apply one of the many procedures currently in use. Over the past decade and a half, many approaches for studying differential item functioning have been developed (see Clauser & Mazor, 1998 for an overview of statistical procedures for identifying DIF). Wainer (1993, pg. 123) divides these into two categories of statistically rigorous procedures that he called empirically-based and model-based. Empirically-based methods include the Mantel-Haenszel chi-square (Holland & Thayer, 1993), the standardization procedure (Dorans & Kullick, 1986) and logistic regression (Swaminathan & Rogers, 1990). With the exception of logistic regression, these are based on the analysis of contingency tables. Model-based indices of DIF such as the likelihood ratio test (Thissen, Steinberg & Wainer, 1988; 1993), Lord's chi-square test (1990), and Raju's Exact Signed Area and H-statistic (1988, 1990) examine the difference between item characteristic curves (ICCs). SIBTEST (Shealy & Stout, 1993),

though similar in some respects to the standardization procedure, would also be considered model-based.

For this study, the Mantel-Haenszel procedure was chosen as representative of popular DIF methods, and results from it were used to judge the problems inherent in the current strategy of using manifest variables. When examining an item for DIF using typical manifest procedures, the groups under consideration are usually comprised of a minority population (e.g., Hispanic examinees or students with disabilities) and a majority population (e.g., white examinees or non-disabled students). In the traditional DIF parlance, these are referred to as the focal and reference groups respectively.

For the Mantel-Haenszel procedure a contingency table (see Figure 1) is constructed for each test score so that examinees in the focal and reference groups can be compared across "homogenous strata" (Meyer, Huynh, & Seaman, 2004, pg. 332). In Figure 1, A_j represents the number of examinees in the reference group who answered an item correctly within the j^{th} level of the matching criterion and B_j represents the number in that group who answered the question incorrectly. C_j and D_j are the corresponding counts for the focal group.

		Score on Item		
		1	0	Total
Group	Reference	A_j	B_j	N_{rj}
	Focal	C_j	D_j	N_{fj}
	Total	M_{1j}	M_{0j}	T_j

FIGURE 1: Contingency table for an item within the j^{th} level of the matching criterion

The Mantel-Haenszel chi-square statistic is then calculated using the following equations. Note that the first equation includes a continuity correction.

$$MH\chi^2 = \frac{\left(\left|\sum_j A_j - \sum_j E(A_j)\right| - \frac{1}{2}\right)^2}{\sum_j \text{var}(A_j)}$$

where the expectation for A_j is calculated using the marginal frequencies and the variance is found using the following equation;

$$Var(A_j) = \frac{N_{rj} N_{fj} M_{1j} M_{0j}}{T_j^2 (T_j - 1)}$$

The resulting statistic, which provides a measure of association between performance on an item and group membership, is distributed approximately as a chi-square with one degree of freedom. An estimate of the constant odds ratio, α_{MH}, can be calculated as an indication of the magnitude of DIF;

$$\alpha_{MH} = \frac{\sum_j A_j D_j \Big/ T_j}{\sum_j B_j C_j \Big/ T_j}$$

Typically, the natural log of this expression is taken so the index is on a symmetric scale with a mean of zero. When that is done, positive values suggest that the reference group is advantaged and the focal group disadvantaged by the item. The opposite holds true for an item with a negative log odds.

There are many advantages to using the Mantel-Haenszel procedure as a representative approach based on manifest comparison groups. Since this is a non-parametric procedure, it may be used with fewer examinees than other approaches. It provides an estimate of the effect size, as well as a statistical test of significance. At the most basic level, the Mantel-Haenszel procedure is also the most intuitively appealing and understandable to those who may have little experience with DIF detection.

A final reason to use the Mantel-Haenszel procedure deals specifically with this research. The model employed in the latent class DIF analysis is in the Rasch (1960) family, for which the sufficient statistic for the latent trait is the total score. In the Rasch model the probability of the n^{th} person getting the i^{th} item correct is represented by:

$$P_{ni} = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)}$$

where θ_n is the ability parameter for person n, and b_i is the item difficulty parameter for item i.

Additionally, the Mantel-Haenszel procedure uses total score to match examinees on ability. Additionally, as Linacre and Wright (1989, pg. 53) point out, "the Rasch and MH approaches are both based on the relative odds of success of the two groups on the suspect item. The difference between the two methods is only in how this relationship is estimated and informed". Since that is the case, by using the Mantel-Haenszel procedure in one part of this research and the Rasch model in the other, some degree of consistency is maintained. It should also be noted that results from the Mantel-Haenszel procedure are generally similar to those from other, more complex ones in certain situations. For example, the Mantel-Haenszel procedure and logistic regression are fairly equal in detecting uniform DIF (Rogers & Swaminathan, 1993; DeAyala, et al, 2002), as are the Mantel-Haenszel procedure and SIBTEST (Roussos & Stout, 1996a).

Mixed Rasch Model

One approach to providing a solution to the problems plaguing the current DIF strategies is to utilize a mixed Rasch model (Rost, 1990). The main thrust of this approach is that the Rasch model can be used to describe "the response behavior of all

persons within a latent class, but that different sets of item parameters hold for the different latent classes" (Rost, 1990, pg. 271). That is, two or more sub-populations of individuals can be identified that are "Rasch scalable" (pg. 271).

Since item response theory and latent class analysis are mixed within this approach, parameters from each need to be included. These parameters are ability parameters under the condition that person n belongs to latent class g (θ_{ng}), item difficulty parameters that are also conditional upon latent class membership (b_{ig}), and latent class proportions (π_g). In latent class analysis it is assumed that the observed responses are independent within latent classes. Therefore, the probability of a correct response by person n to item i can be expressed as a weighted sum of conditional probabilities. That can be given by the following equation:

$$p_{ni} = \sum_g \pi_g \frac{\exp(\theta_{ng} - b_{ig})}{1 + \exp(\theta_{ng} - b_{ig})}$$

It is noteworthy that in this formulation there is not an assumption that an individual belongs to a certain latent class and that only the parameters associated with that class should be applied. Instead the probabilities associated with membership in each class are applied to each person.

Rost's mixed Rasch model is one of several approaches that could have been employed in this research. Among these are the IRT model for different strategies (Mislevy & Verhelst, 1990), and the loglinear models of Kelderman and Macready (1990). Mislevy and Verhelst (1990) posited that standard IRT models are not satisfactory when examinees employ different strategies for solving problems. Though this model is similar in some respects to the mixed Rasch model, it is quite different in

that "[s]ubstantive theory associates the observable features of items with the probability of success for members of each strategy class" (Mislevy & Verhelst, 1990, pg. 198). Since this research can best be described as exploratory in nature with no substantive theory regarding item features, Mislevy and Verhelst's model seems inappropriate for the present study. The use of the mixed Rasch to investigate DIF corresponds to one of the cases of the loglinear models of Kelderman and Macready (1990). One of their models includes terms for the interactions between the latent variable and the items. Comparing this model with a null model is equivalent to testing all of the items for DIF. For this research the mixed Rasch model was chosen over the analogous loglinear model because it is rooted in IRT. Since this is the case, differential item functioning can be determined by looking at the differences between the item difficulties for the latent classes. For those accustomed to IRT parameters, this should directly yield a much more intuitive result regarding the magnitude of the DIF. In addition, as Kelderman and Macready acknowledged, parameter estimation in their models may be difficult when there are large numbers of variables as there would be on educational tests.

Bayesian Inference

When estimating parameters in IRT, a marginal maximum likelihood solution can be found using the Expectation-Maximization (EM) algorithm (Bock & Aiken, 1981). Although this approach is often useful, it has drawbacks. When models are extremely complex this approach can be cumbersome, for example, due to the calculation of derivatives of non-linear functions. Patz and Junker (1999, pg. 146) note that, "In contrast to this two-stage E-M approach, Markov chain Monte Carlo (MCMC) methods treat item and subject parameters at the same time; this allows us to incorporate standard

errors of item estimates into trait inferences, and vice versa." To fit complex models and incorporate uncertainty, researchers have begun using MCMC algorithms (Gelman, Carlin, Stern & Rubin, 1995; Gilks, Richardson & Spiegelhalter, 1996). These procedures simulate random samples from a theoretical distribution and then use those samples in making inferences about the features of that theoretical distribution.

The goal of Bayesian modeling is to define a posterior distribution as opposed to arrive at a point estimate. One applies Bayes theorem, which states that the posterior distribution of a parameter is equal to the product of the prior density and the likelihood of that parameter divided by the marginal probability of the observed variables integrated (or summed in the discrete case) over the unknown parameters. When it is not possible to obtain an analytical solution in that manner, one is able to utilize MCMC estimation and uncover the posterior distribution by taking a large number of draws from that distribution.

To begin the process of constructing a Markov chain, start values for all parameters are given; this is know as initializing the chain. Subsequent values for these parameters are drawn using one or a combination of sampling schemes. Popular examples of these techniques are Gibbs (see Casella & George, 1992 for an explanation), Metropolis (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) and Metropolis-Hastings (Hasting, 1970) sampling. Assuming certain regularity conditions hold (see Tierney, 1994) these chains of values for each parameter will eventually converge to a target (or stationary) distribution, which is the posterior distribution of the parameters in the model. The draws taken before these chains reach stationarity are known as 'burn-in' and are typically discarded. After stationarity is attained, the remaining draws will be

distributed like draws from the posterior distribution, and parameter estimates can be obtained by sampling from that distribution.

In this research, the WINBUGS software (Spiegelhalter, et al., 2000) was used in the analysis of the models using MCMC. In addition to the aforementioned benefits, this strategy is also advantageous because it is so flexible. Since everything is treated as a random quantity, it is simple to incorporate additional information about students, items, or both. For example, it may be appropriate to start with the prior distributions for both the ability and item difficulty parameters being normal (0,1). At a later point the difficulty parameters could be modeled in terms of some elementary components contributing to the difficulty. The same can be done for the ability parameters, latent class proportions, and conditional probabilities. Because of this flexibility, it was a relatively simple matter to find the differences between the difficulty parameters for the two latent classes and the proportion of each manifest group within each latent class. It was also possible to examine the latent ability distributions using this strategy, rather than the distributions for the manifest groups as with other strategies.

Chapter 3: Methodology

This research was multi-faceted. First, the case needed to be made for using a latent class, rather than a manifest group approach for studying differential item functioning. To highlight the inadequacies of manifest approaches to DIF detection the Mantel-Haenszel procedure was used as a representative approach and simulated data with varying amounts of DIF were analyzed. Since the characteristics of these data were known, there existed a standard against which to judge the efficacy of the currently employed types of procedures. Second, the mixed Rasch model was used on a subset of these simulated data to investigate the strengths and weaknesses of that approach. Next, a strategy for applying a latent class approach needed to be delineated. This included the development of a series of protocols and recommendations for use. Finally, this strategy was applied to real data from an assessment of English Language proficiency that was being field-tested at the time the data was collected. This test was chosen because a relatively large number of covariates were available along with the item responses.

Making the Case Using the Mantel-Haenszel Procedure

The first part of this research provides a justification for thinking about DIF in a different manner. The impact of the amount of overlap between latent classes and manifest groups had on each of the following was examined:

- Number of items correctly identified as having DIF or the power to detect differential functioning;
- Magnitude of the DIF which provides a measure of meaningfulness regarding the differential functioning;

- Type 1 error rate or the number of items falsely identified as functioning differentially.

This research built upon the recent work done by DeAyala, et al. (2002), but the treatment of the overlap between latent classes and manifest groups was more methodical than in previous research. The expectation is that this investigation will delineate conditions under which the lack of overlap becomes problematic and highlight the extent of problems faced. This may impact the decisions researchers make for DIF studies in terms of sample size, and the magnitude of the DIF considered meaningful enough to prohibit an item from remaining on an operational test.

Factors

Data were simulated for a fixed length test of 20 items, mimicking the length of the subtests used in a study with a similar conception of DIF (DeAyala, et al., 2002). This test length was also practical given that these data were examined using WINBUGS, a flexible but notoriously time consuming computer software, in other parts of this research. These data were simulated with either 500 or 2000 total examinees, sample sizes that are consistent with those used in other studies of differential item functioning (Narayanan & Swaminathan, 1994, 1996; Penfield, 2001). In addition to sample size, five other factors were manipulated. They were the:

- manifest proportions,
- overlap between the manifest groups and the latent classes,
- number of items exhibiting DIF,
- effect size of the DIF,
- ability distributions within the latent classes.

In this study, two conditions were considered with regard to the manifest proportions – a 50/50 examinee split and an 80/20 split. The first of these could represent gender differences in a population, while we may think of the latter as simulating a condition in which differences in item function exists between the majority and minority groups. Within these manifest situations two latent classes were examined under five different overlap conditions – 100%, 90%, 80%, 70%, and 60%. In this research 100% overlap refered to a condition in which one latent class was comprised entirely of examinees from a single manifest group and the other overlap conditions refers to situations in which the latent classes were increasingly more heterogeneous with regard to the manifest groups. The resulting numbers of examinees within the classes for the condition in which there were 2000 examinees are shown in Table 1. For 500 examinees these values were divided by four.

TABLE 1
Breakdown of numbers of examinees within each latent class for 2000 total examinees

		50/50 Manifest Split		80/20 Manifest Split	
% Overlap		LC1	LC2	LC1	LC2
100	Manifest Group 1	1000	0	1600	0
	Manifest Group 2	0	1000	0	400
90	Manifest Group 1	900	100	1440	160
	Manifest Group 2	100	900	40	360
80	Manifest Group 1	800	200	1280	320
	Manifest Group 2	200	800	80	320
70	Manifest Group 1	700	300	1120	480
	Manifest Group 2	300	700	120	280
60	Manifest Group 1	600	400	960	640
	Manifest Group 2	400	600	160	240

The factor for the number of items with DIF had three levels with 2, 6 or 10 items functioning differentially. Since there were twenty items, this resulted in 10%, 30% or 50% of the items having DIF. While many results for actual tests (Hambleton & Rogers, 1989; Raju, Bode & Larsen, 1989) with large amounts of DIF report between 15% and 30% of the items exhibiting DIF, it was necessary to span a wider range in this case. Looking at differential item functioning using a latent class approach should maximize the differences between groups. If that were the case, one would expect to encounter more items with DIF than we normally would using manifest grouping variables; hence the larger percentages were appropriate. The effect size factor also consisted of three levels based on the difference between the item difficulty parameters in the two latent classes. These differences were $\Delta b = 0.4$, $\Delta b = 0.8$, and $\Delta b = 1.2$. The smallest amount of DIF ($\Delta b = 0.4$) was consistent with magnitudes used in other simulation studies (Clauser, Mazor, & Hambleton, 1993; Penfield, 2001). The larger differences in item difficulty parameters were included because it was theorized that larger **latent** differences might result in the smaller amounts of **manifest** DIF typically found. Item difficulty parameters were evenly distributed between –2.0 and 2.0 for the first class. Adding and/or subtracting Δb for the specified items generated the corresponding parameters for the second class. The factors for the number of items and the effect size were not fully crossed; instead they were manipulated as shown in Table 2. The items altered to incorporate DIF were chosen to yield item difficulties that were at or around zero.

To generate data for the 2000 (or 500) examinees on these 20 items, simulees were assigned to a latent class and a manifest group according to the design discussed above. Values for the ability parameters were generated for simulees in each latent class

by randomly sampling from a standard normal distribution. In the case where the ability distributions differed between the latent classes, theta values for those in the second class came from a normal distribution with a mean of -1. This is consistent with other DIF studies (Hambleton & Rogers, 1989; Narayanan & Swaminathan, 1996) in which manifest groups have different means.

TABLE 2
Increments added on to item difficulties for the first latent class b[i,1] to create the difficulties for the second class b[i,2].

	b in LC1	Increments added for b in LC2			
		Small DIF	Medium DIF	Large DIF	Mixed
2 items	b[3,1]	+ 0.40	+0.80	+1.20	NA
	b[5,1]	-0.40	-0.80	-1.20	NA
6 items	b[3,1]	+0.40	+0.80	+1.20	+1.20
	b[4,1]	+0.40	+0.80	+1.20	+0.80
	b[5,1]	+0.40	+0.80	+1.20	+0.40
	b[6,1]	-0.40	-0.80	-1.20	-0.40
	b[7,1]	-0.40	-0.80	-1.20	-0.80
	b[8,1]	-0.40	-0.80	-1.20	-1.20
10 items	b[1,1]	+0.40	+0.80	+1.20	+0.40
	b[2,1]	+0.40	+0.80	+1.20	+0.80
	b[3,1]	+0.40	+0.80	+1.20	+1.20
	b[4,1]	+0.40	+0.80	+1.20	+0.80
	b[5,1]	+0.40	+0.80	+1.20	+0.40
	b[7,1]	-0.40	-0.80	-1.20	-0.40
	b[8,1]	-0.40	-0.80	-1.20	-0.80
	b[9,1]	-0.40	-0.80	-1.20	-1.20
	b[10,1]	-0.40	-0.80	-1.20	-0.80
	b[11,1]	-0.40	-0.80	-1.20	-0.40

Using the Mixed Rasch Model on Simulated Data

In addition to using the Mantel-Haenszel technique on the simulated data, the mixed Rasch model was also employed to detect DIF to verify that the model under consideration in this research is sound. For purposes of practicality, a subset of the simulated data was used with the mixed Rasch model. Specifically, the cases with 500 and 2000 examinees with a 50/50 manifest split and differences in the ability distributions of the latent classes under the 60% through 90% overlap conditions were examined. The data with 500 examinees and differences in the mean abilities of the latent classes represent a challenge in terms of the estimation of the parameters. Therefore, if the mixed Rasch model can effectively detect DIF for these data, there is evidence of its tenability under a variety of conditions.

For each condition the differences in item difficulties between classes, the means of the latent ability distributions, and the percentages of examinees from each manifest group in each latent class were determined using Markov Chain Monte Carlo (MCMC) estimation with the WINBUGS computer program (Spiegelhalter, Thomas & Best, 2000). In order to estimate the model parameters the following prior distributions were used.

Item difficulties (i) within classes (g): $b[i,g] \sim \text{Normal}(0,1)$

Ability distributions within classes: $\theta[n,g] \sim \text{Normal}(\mu_g,1)$

Means of the ability distributions within classes: $\mu[g] \sim \text{Normal}(0,1)$

Item responses for examinees on items: $x[n,i] \sim \text{Bernoulli}(P[n,i])$

These were consistent with the prior distributions recommended by other researchers (Bolt, Cohen, & Wollack, 2002; Wollack, Cohen, & Wells, 2003) to ensure convergence with comparable mixture models.

Further constraints were placed on the model in terms of the item difficulties within classes summing to zero. In addition to solving the indeterminacy issue, these constraints have the effect of putting both sets of item parameters on the same scale in the sense that if there is no DIF they are the same within estimation error, and if there is DIF, the items parameters being centered around zero in both groups makes the DIF average to zero. In order to make the item parameters center around zero, the item difficulties were estimated for the first J-1 items (where J is the total number of items) and the item difficulty for the Jth item was then defined as the negative sum of the other items within a given class. Within BUGS, DIF was then defined as the difference between item difficulties across latent classes, the posterior distributions were monitored and those items that had 95% posterior credibility intervals containing zero were identified as functioning differentially.

When running MCMC using BUGS, one of the first steps was to determine the number of iterations needed to reach convergence. These are known as the burn-in, and are typically discarded so that theoretically only draws from the posterior distribution are used. Convergence of the MCMC algorithms can be assessed in a number of ways. In this study, the following graphical methods of gauging convergence were utilized.

- Time series plots – when convergence is reached, these plot shows random sampling within the same part of the same space for all chains
- Plots of the auto-correlation function – these plots show the relationship of draws for a variable from one cycle to the next and in doing so they indirectly assist in assessing convergence by providing information regarding why the chain(s) may be traveling slowly across the sample space
- Brooks-Gelman-Rubin (BGR) diagnostic plots – comparing between- and within-chain variation for chains with divergent starting values

After convergence is achieved, enough iterations should be run to have confidence in the inferences made about the posterior distributions. One method of assessing whether

enough iterations have been completed is to examine the density plots. Once the posterior distribution is sampled fully one would expect to see smooth curves like those shown in Figure 2. A rule of thumb often applied is that the simulations should be run until the ratio of the standard deviation to the Monte Carlo standard error of the mean (MC error) is less than 0.05 (Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D., 2003). Both of these methodologies were employed in this research.

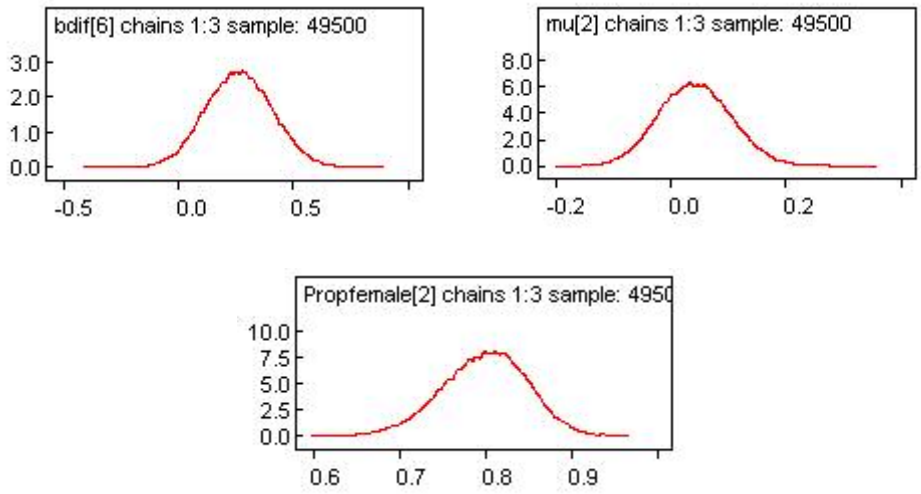


FIGURE 2: Representative density plots for the difference in item difficulties between classes (bdif), the latent ability distributions, and the proportions of males and females within latent classes.

Development of Protocols and Their Use on Data from an Operational Test

The final part of this research started with the development of the protocols for a latent class approach to DIF utilizing MCMC. These protocols are a step-by-step procedure for examining differential item function along with recommendations or cautions for doing DIF analyses within BUGS.

After these simulation studies were completed and the protocols had been developed, this technique was applied to a data set for a reading test of English Language

Acquisition. This test has been chosen for two reasons. First, it was an assessment that was still being field-tested, so all items functioning differentially would not have been removed. Though it is doubtful that any test, even an operational one, can be DIF-free, the fact that more of these items will function differentially makes this a more interesting test to study. The second reason this test was chosen was due to the relatively large amount of background data being collected. In this case, the following data of interest was collected for all students in addition to the typical demographic variables (age, race, ethnicity):

- Born in the United States – a dichotomous variable indicating whether the English language learner was born in the United States or another country
- Years receiving ESL instruction – a continuous variable
- Grade of the student – either 3rd, 4th, or 5th
- Home language

These background variables were then used as a series of categorical and continuous covariates. In should be noted that because this was a test of English proficiency for students who do not speak English as a first language, the 'focal' and 'reference' groups chosen were Asian and Hispanic speakers. Evidence of validity in this case would be that the items on the test do not function differentially for students speaking different languages.

In this research, one-, two- and three-class models were checked using a technique employing "shadow" data sets created during the MCMC iterations which enable a comparison of the observed data to the posterior predictive distribution (Gelman, et al., 1995). Assuming the model used fits, the distribution of the shadow data is a null distribution from which the actual data are plausible draws. To facilitate this comparison of the observed and shadow data, a test quantity was defined to measure the discrepancy

between the simulated values (also called shadow data) and the observed data. In this case, the mean square error was the test quantity of interest. Shadow data was generated from the model (the posterior predictive distribution), one in which item responses were modeled as a Bernoulli random variable with a probability of success that was a function of latent class membership, ability, and item difficulty. The squared differences between the observed data and the expectation, and between the shadow data and the expectation, were calculated. A measure of person fit was then calculated for both the real and shadow examinees by taking the square root of the mean of those squared differences. A count was then made of the number of times the observed data was worse than the shadow data. If the model was correct, the observed data should only randomly fit worse than the shadow data. Therefore, if the proportion of times the observed fits worse than the shadow is statistically different from 0.5, there is evidence that the model does not fit the data.

Chapter 4: Results of Simulations with Mantel-Haenszel

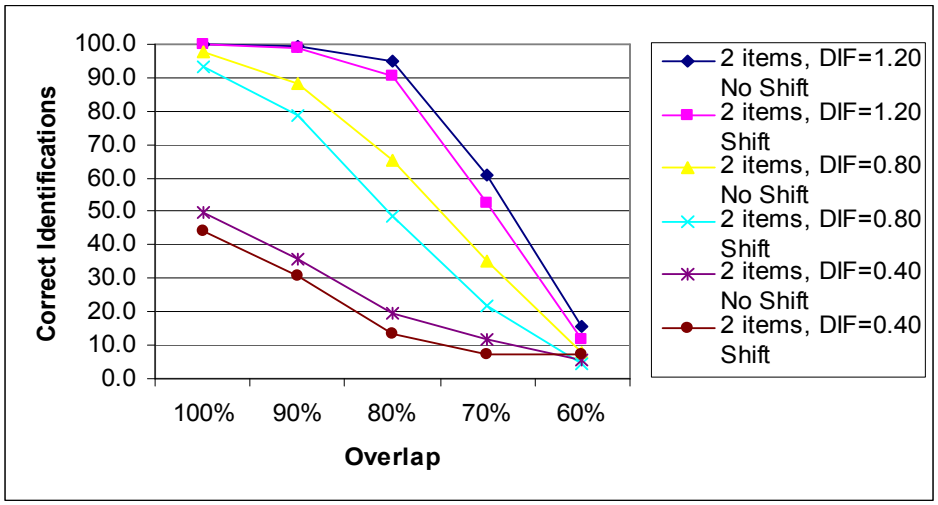
There are three broad categories of results from analysis of the simulated data using the Mantel-Haenszel procedure. The first category deals with the number of correct identifications made for the items that function differentially, or the power of the procedure to identify items with DIF. The second group of results deals with the meaningfulness of the DIF as measured by the ln(odds). Finally, the number of false positive identification of items without DIF is examined.

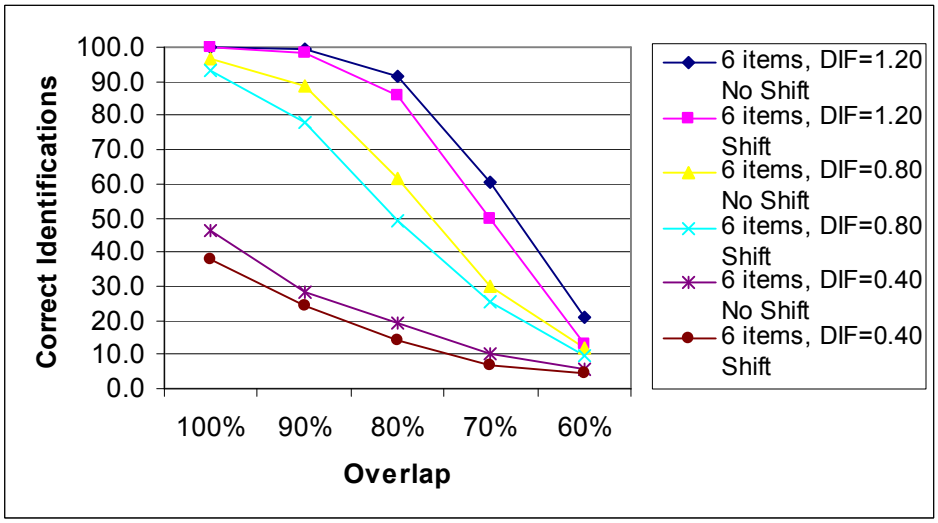
Power to identify DIF items

This research provides the following five insights into the number of correct identifications of items that contain DIF.

1. The number of correct identifications decreases as the amount of overlap between the manifest groups and latent classes decreases.
2. When the ability distributions of the latent classes are the same there are more correct identifications than when they differ.
3. The number of correct identifications decreases with increased contamination of the matching criterion by items with DIF.
4. More correct identifications are made when the sample size is larger.
5. More correct identifications are made in the 50/50 condition than in the 80/20 condition.

Some of these results are intuitive and are supported by other research in DIF. Each will be discussed in more depth below. To facilitate the discussion, Figures 3 through 6 are provided (also Appendix A) showing the correct number of identifications as a function of the percent overlap, effect size, number of DIF items, differences in the ability distributions, sample size, and manifest proportions. Each data point in the graphs of those figures represents the average for the DIF items under a particular set of conditions.





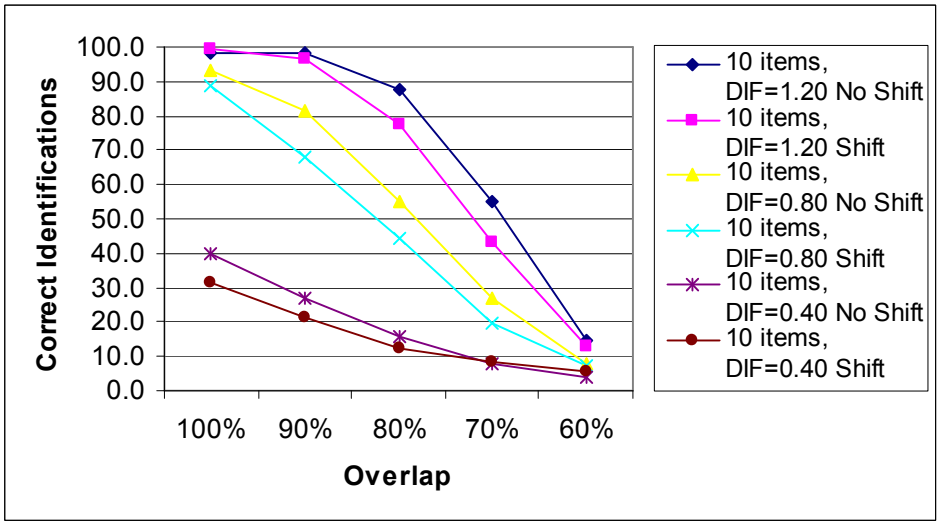


FIGURE 3: Correct identifications for 50/50 split with 500 examinees

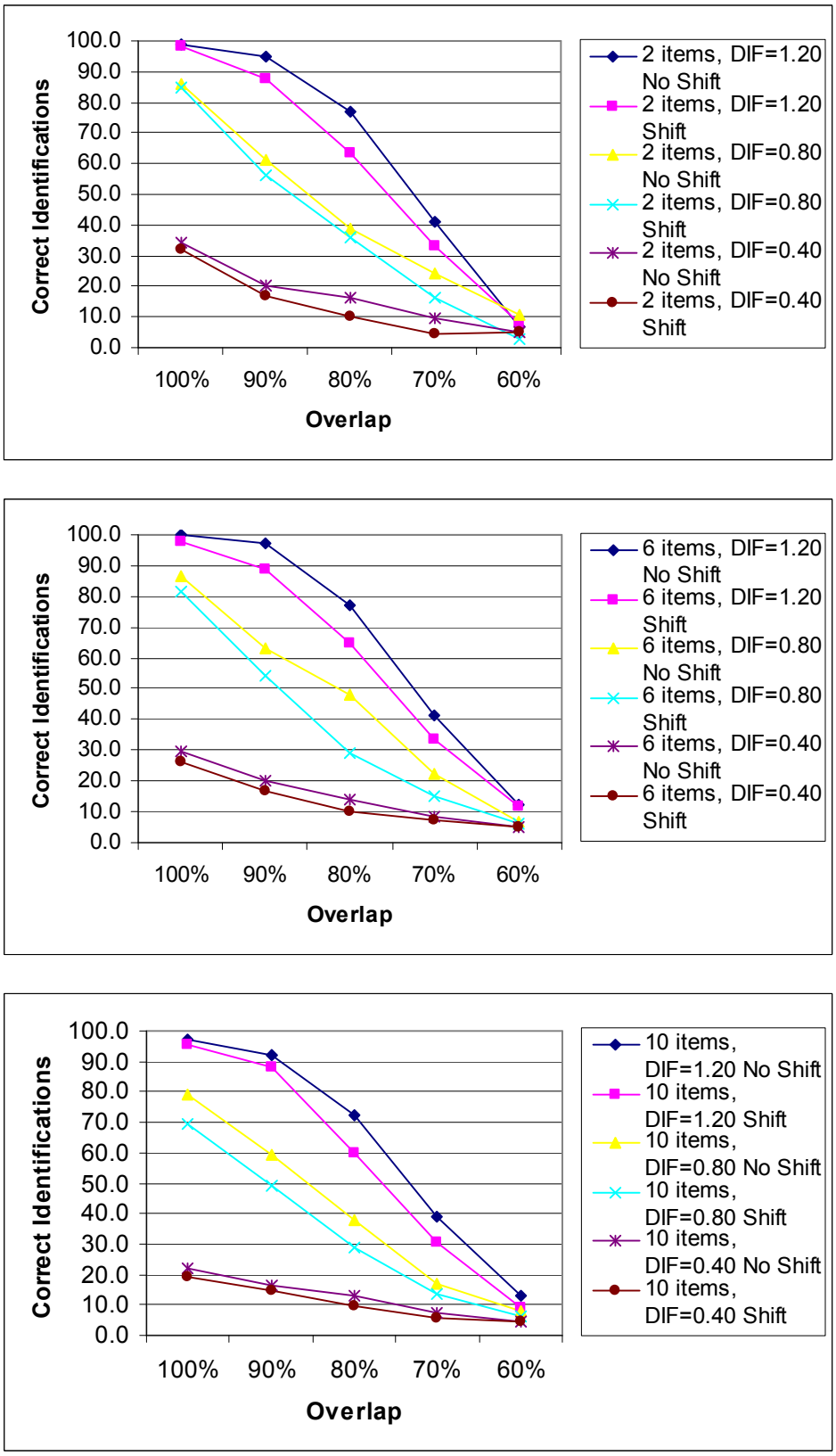
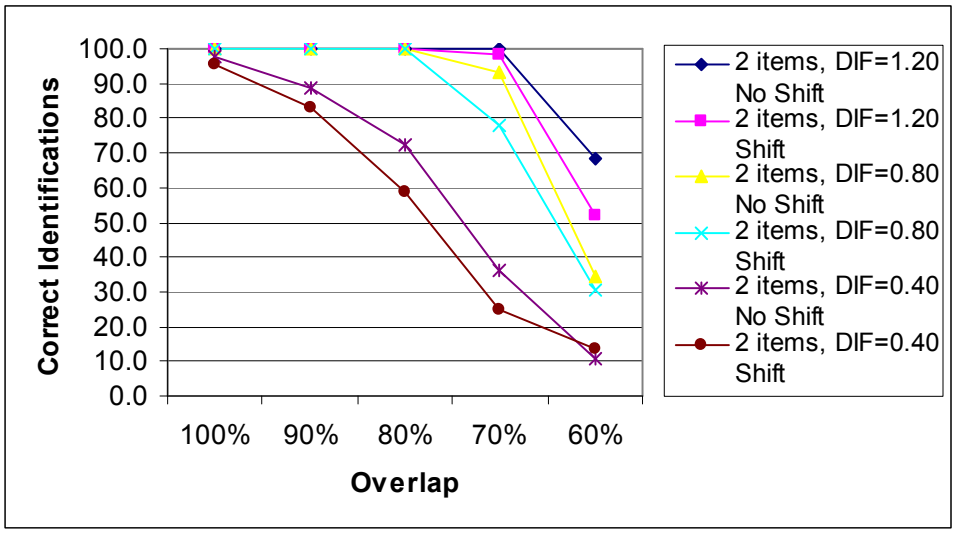
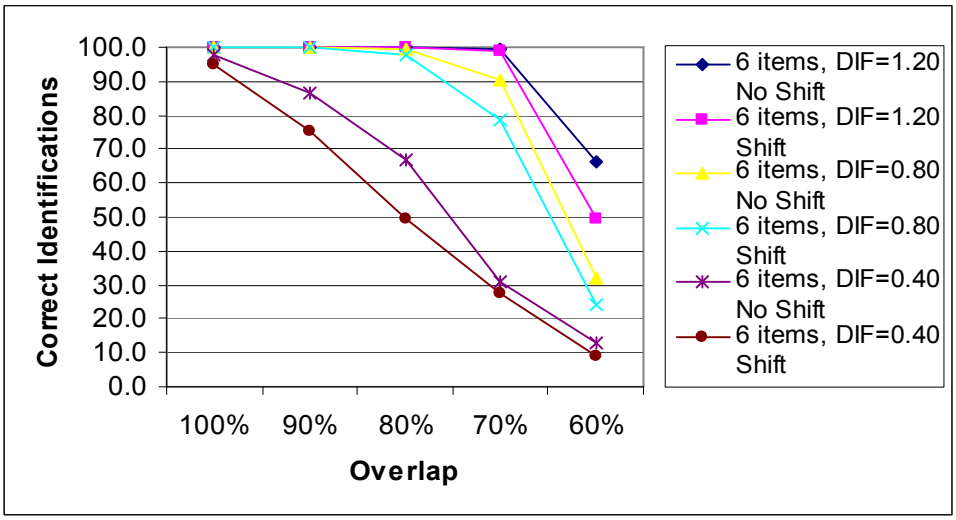


FIGURE 4: Correct identifications for 80/20 split with 500 examinees





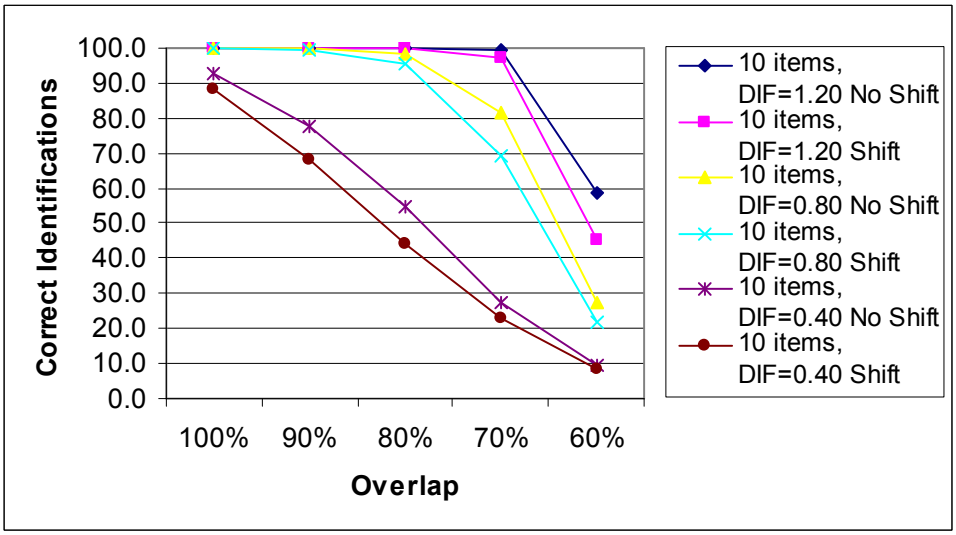


FIGURE 5: Correct identification for 50/50 split with 2000 examinees

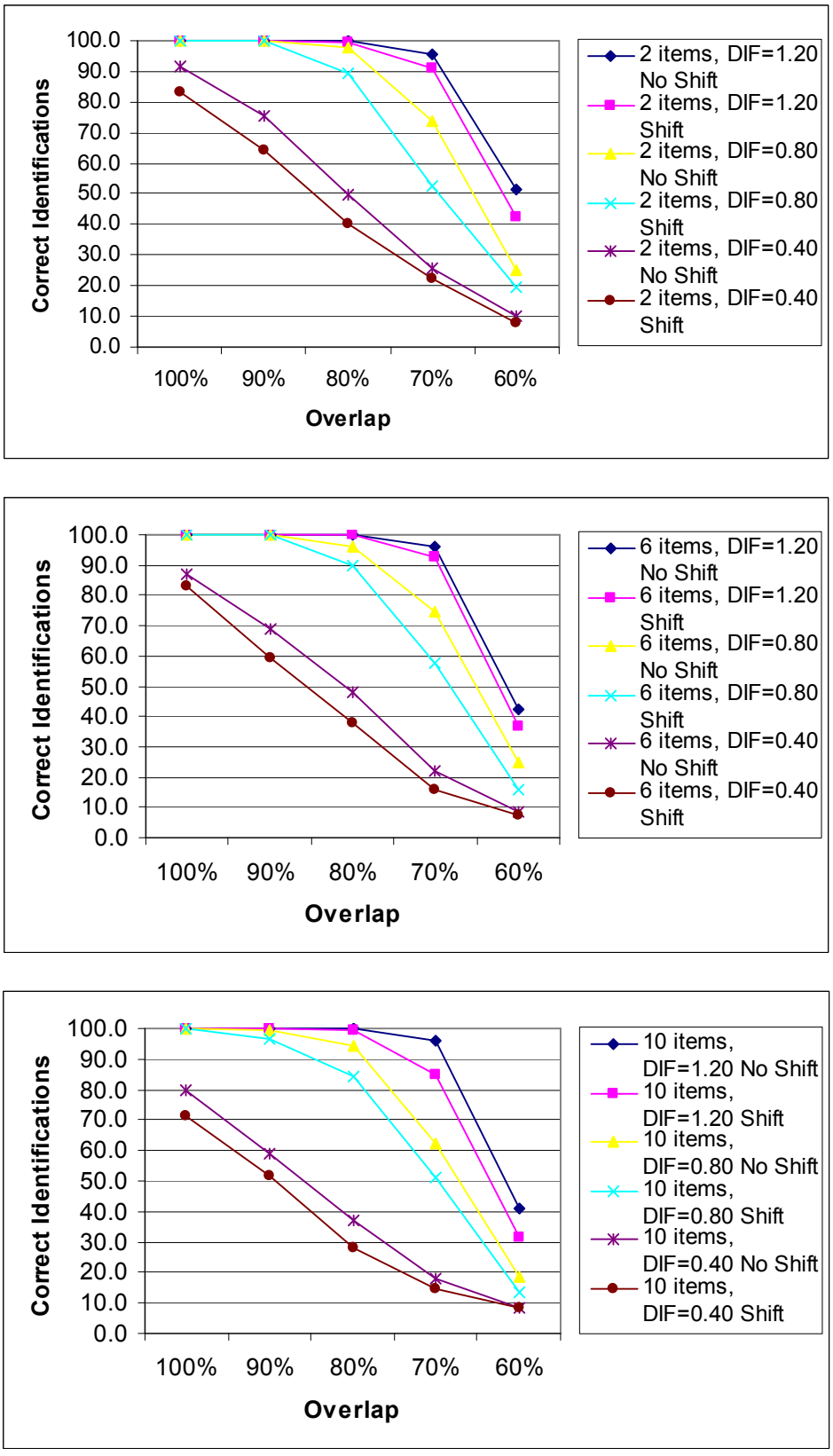


FIGURE 6: Correct identifications for 80/20 split with 2000 examinees

The downward slopes of the curves in the graphs in Figures 3 through 6 show the impact of the decreasing overlap of latent and manifest groups on power. It is not surprising that it is more difficult to correctly identify items that function differentially as a result of membership in a latent class when the amount of overlap between those latent classes and the manifest groups decreases. As the amount of overlap decreases the membership of each latent class becomes more of a mixture of the manifest groups. That is, an overlap of 100% between latent class and gender would mean one class was entirely made up of males and the other females. On the other end of the spectrum, with overlap of 60%, both males and females will be well represented in each of the latent classes, making it increasingly difficult to see differences between those classes. Using the standard set by Cohen (1988), that power is considered to be sufficient (at the 0.05 significance level) when it is above 0.80, it is apparent that lack of overlap causes problems. As can be seen in Table 3, when the magnitude of the differential functioning is small ($\Delta b=0.40$) there is not sufficient power to see DIF with 2000 examinees when any overlap exists regardless of the percentage of examinees in the manifest groups. Furthermore, there is never sufficient power to identify items functioning differentially when the magnitude of the DIF and sample size are small. As the magnitude of the DIF gets larger there is power to see it even when there is some overlap. However, even in the most advantageous situation with regard to power, large DIF, a 50/50 manifest split, and 2000 examinees, sufficient power exists only down to an overlap of 70%.

TABLE 3
The overlap necessary to achieve power = .80 (at the 0.05 level) at various magnitudes of DIF when the manifest groups are split 50/50 and 80/20 for small to moderate DIF contamination.

	2000 Examinees		500 Examinees	
	DIF Magnitude	Overlap Needed	DIF Magnitude	Overlap Needed
50/50 Manifest Split	1.20	0.70	1.20	0.80
	0.80	0.80	0.80	0.90
	0.40	1.00	0.40	Never sufficient
80/20 Manifest Split	1.20	0.70	1.20	0.90
	0.80	0.80	0.80	1.00
	0.40	1.00	0.40	Never sufficient

The presence of a shift in the ability distributions of the latent classes also decreases the power to see differential item function. Once again, examination of Figures 3 through 6 clearly illustrates that all curves for the shifted condition are below the corresponding curves for the condition in which there is no difference between the ability distributions. This trend appears to be more problematic at the non-extreme overlap conditions (ie. 70%, 80%, and 90%). At these mid-range conditions the number of misidentifications of items with DIF increases by between 5 and 15 per 100 replications. This finding, that a difference in the ability distributions impact power, is consistent with previous research (Mazor, Clauser & Hambleton, 1992; Narayanan & Swaminathan, 1994, 1996).

Results of this study indicate that as the percentage of items functioning differentially increases the power to detect DIF generally decreases. This is evidenced in the trios of graphs in Figures 3 through 6. In each figure, moving from the top graph with 10% of the items exhibiting DIF to the bottom with 50% DIF, generally shows fewer and fewer correct identifications. The only glaring exceptions to this trend appear under the

80/20 condition with DIF=0.80 and 500 examinees. This general result of decreasing power with greater contamination is consistent with outcomes of several previous studies (Clauser, Mazor & Hambleton, 1993; Shealy & Stout, 1993; Fidalgo, Mellenbergh & Muniz, 2000). Since the total score is used as the matching criterion for the Mantel-Haenszel procedure, any contamination of this criterion will impair the effectiveness of matching examinees. This will, in turn, negatively impact the power to detect which items are functioning differentially.

Samples of the graphs in Figures 3 through 6 have been summarized in Figure 7 to illustrate the impact of sample size and the mixing proportions of the manifest groups. The case shown, in which the difference between the item parameters is small ($\Delta b=0.40$), is representative of all. It is apparent from these composite graphs that having a larger number of examinees results in more correct identifications. This is also consistent with the findings of previous studies (Shealy & Stout, 1993; Narayanan & Swaminathan, 1996).

It is also clear from Figure 7 that there is less power with an 80/20 split of the manifest groups than with the equal split. In effect, this is also a sample size issue in that the size of the smaller group (the focal group in this case) will impact detection rates. As that group gets larger, as it does moving from an 80/20 split to a 50/50 split, there is more power to identify items functioning differentially. This result was also noted by Narayanan and Swaminathan (1996).

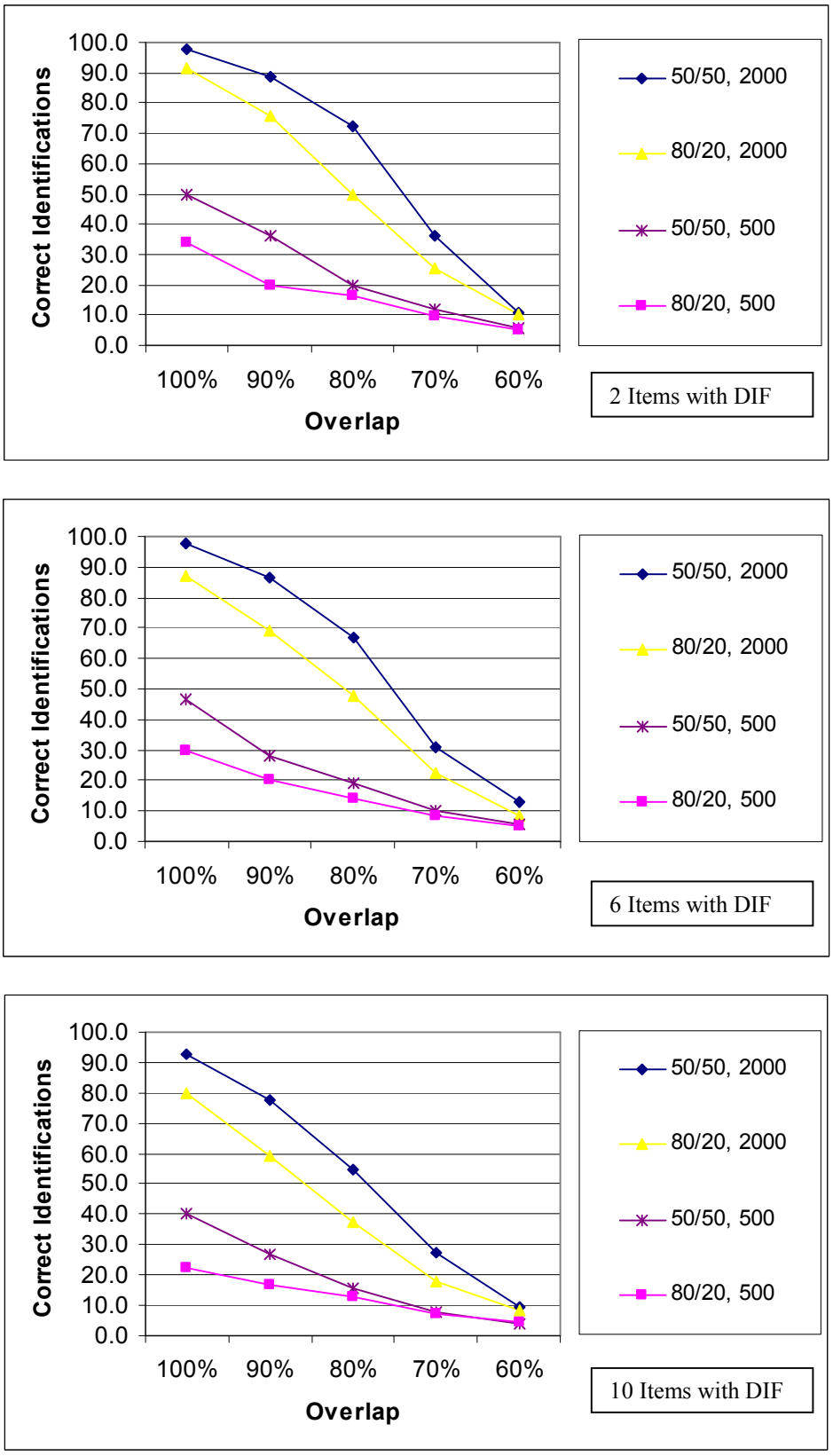
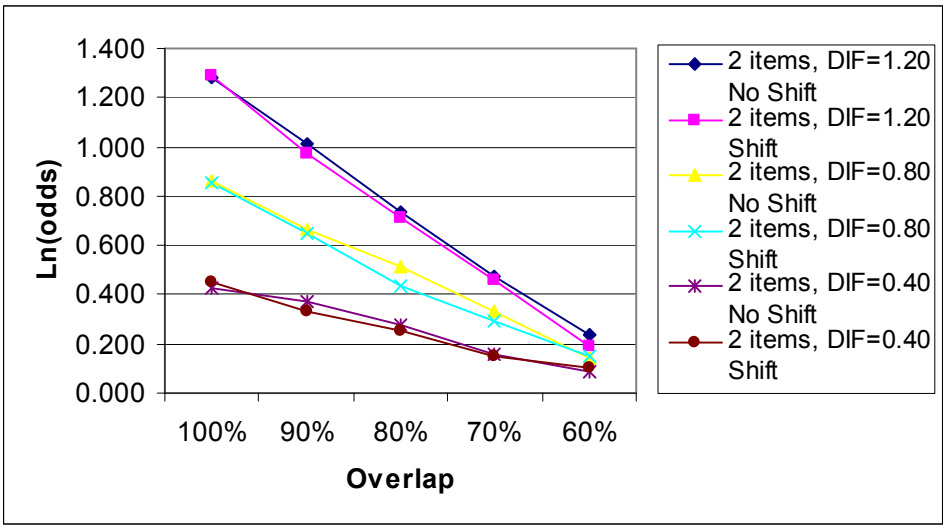


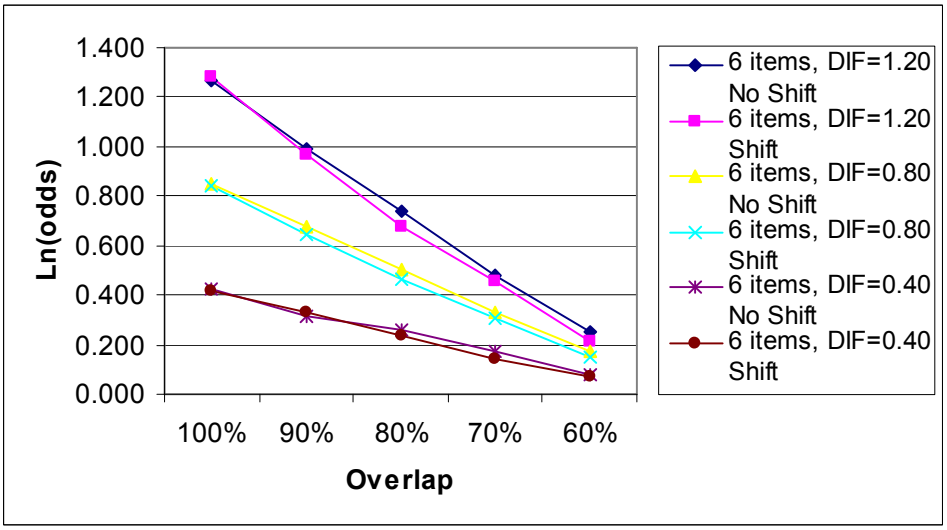
FIGURE 7: Correct Identifications for conditions in which 2, 6 or 10 items contain DIF=0.4

Ln(odds)

While the first step in any DIF detection strategy is to determine which items exhibit DIF, an equally important next step is to decide whether that differential function is large enough to be meaningful. For the Mantel-Haenszel procedure this is typically done by examining the ln(odds) for the item. ETS (Dorans & Holland, 1993) employs a linear transformation (-2.35*ln(odds)) to define their own measure of DIF. Values of this are then used to classify items into one of three categories for the purpose of choosing the items to retain for use on operational tests. Items in category A are those with "negligible or nonsignificant DIF" (Zieky, 1993, pg. 342). Category B includes items with slight to moderate DIF that may be used on test forms with the caveat that items with smaller absolute values of DIF are preferred over those with larger values. Items in category C are generally not used for operational tests since those contain moderate to large amounts of DIF.

Results from the current study have been summarized in Figures 8 through 11. These show the absolute values of the ln(odds) as a function of the percent overlap, effect size, number of DIF items, difference in the ability distributions, sample size, and manifest proportions. Each data point in the graphs of those figures represents the average absolute value of the ln(odds) for the DIF items under a particular set of conditions.





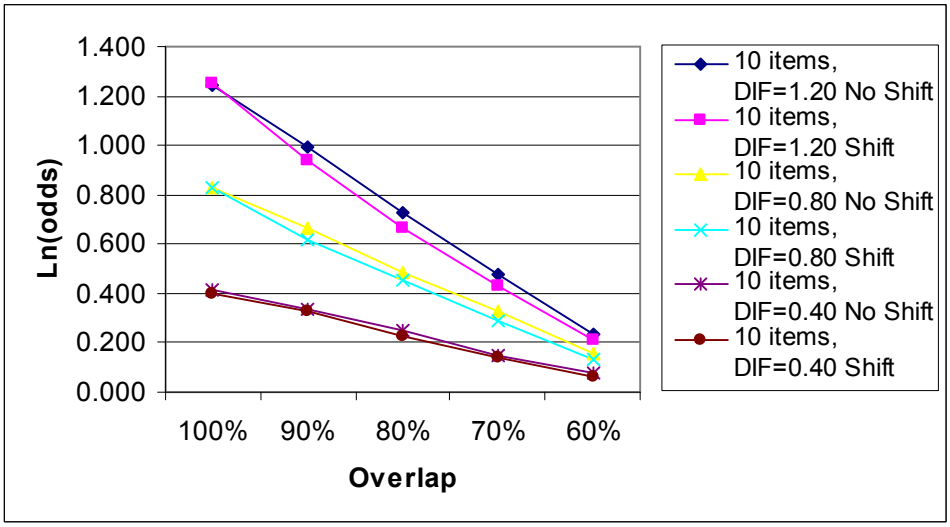
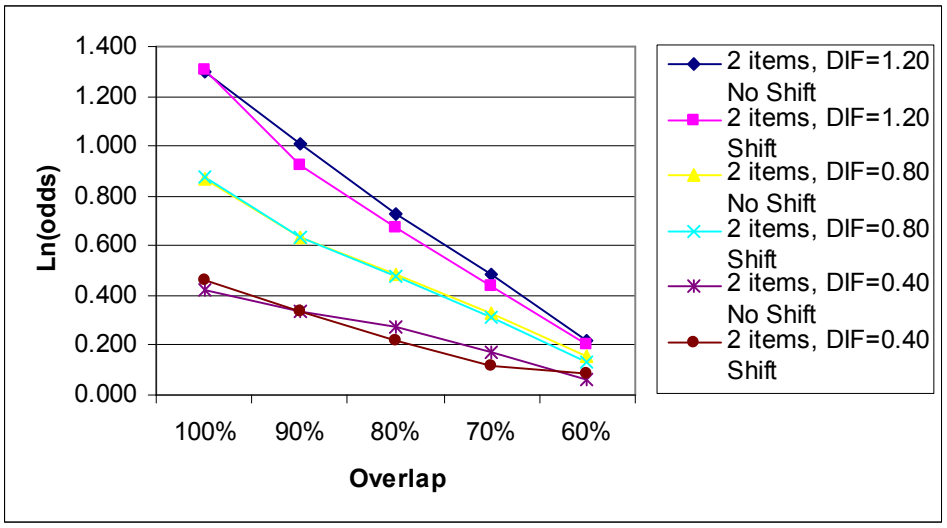
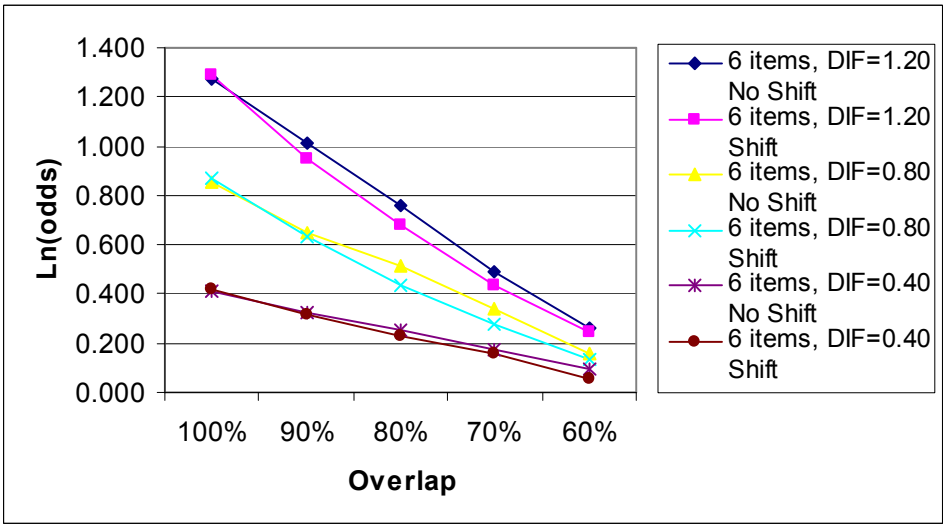


FIGURE 8: Ln(odds) for 50/50 Split with 500 examinees





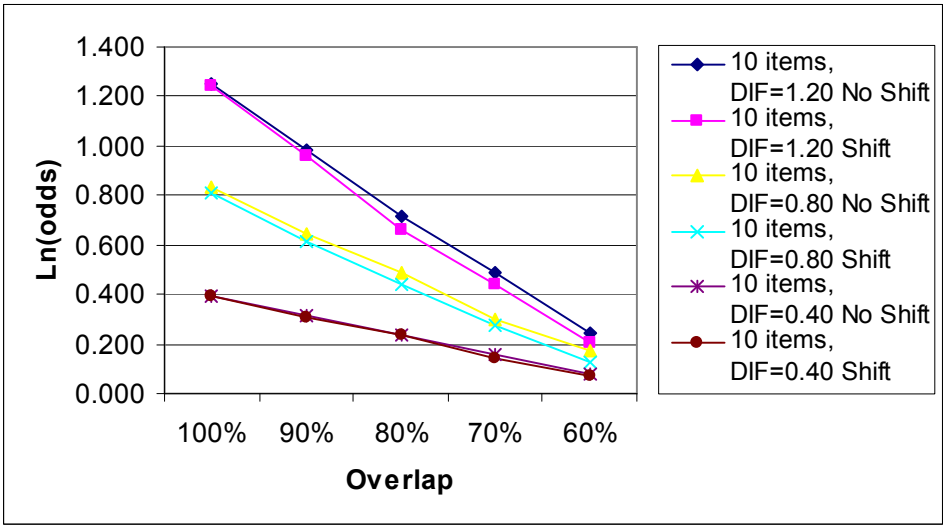
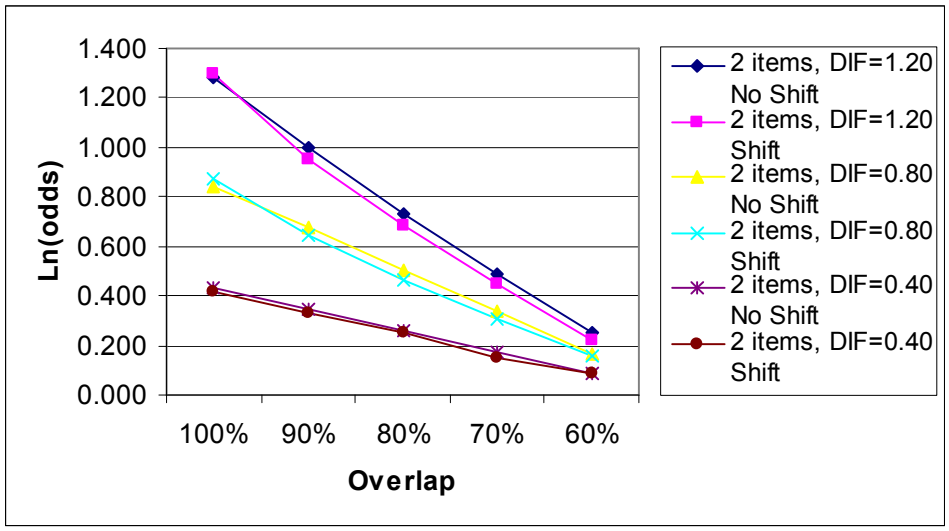
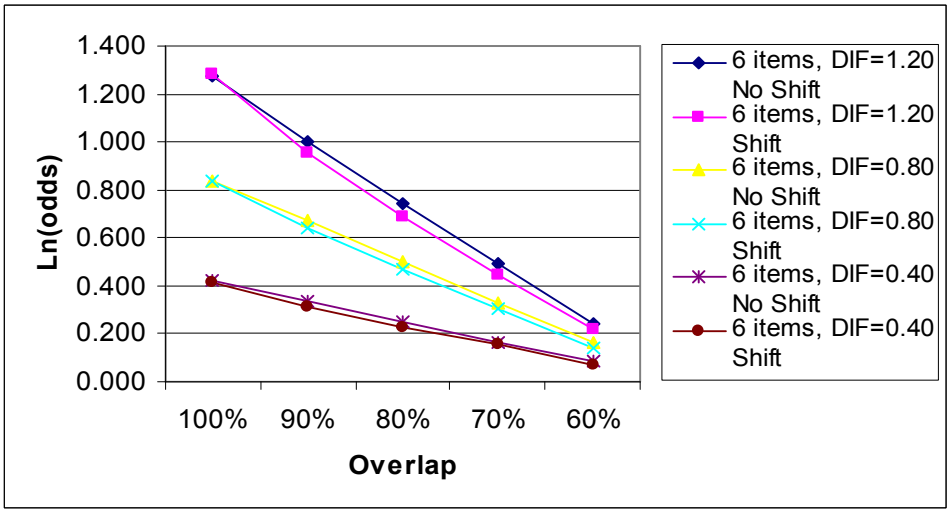


FIGURE 9: Ln(odds) for 80/20 Split with 500 examinees





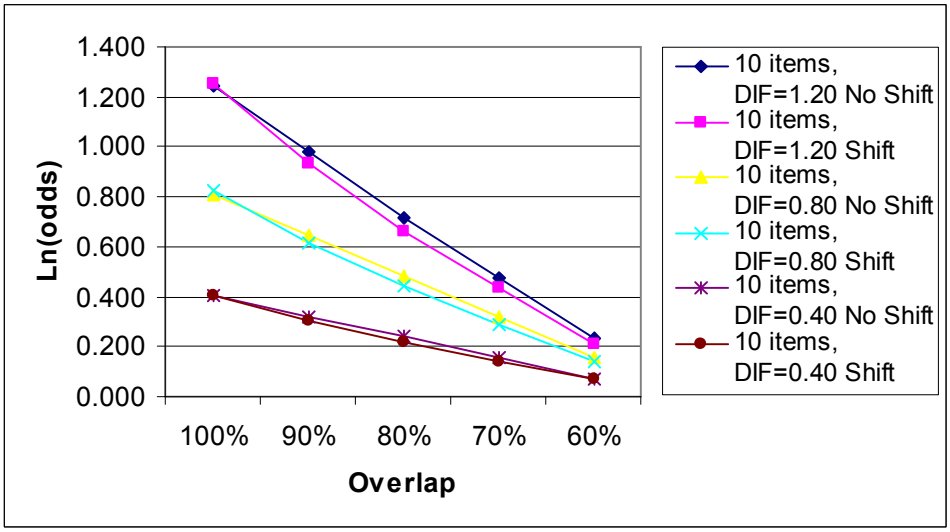
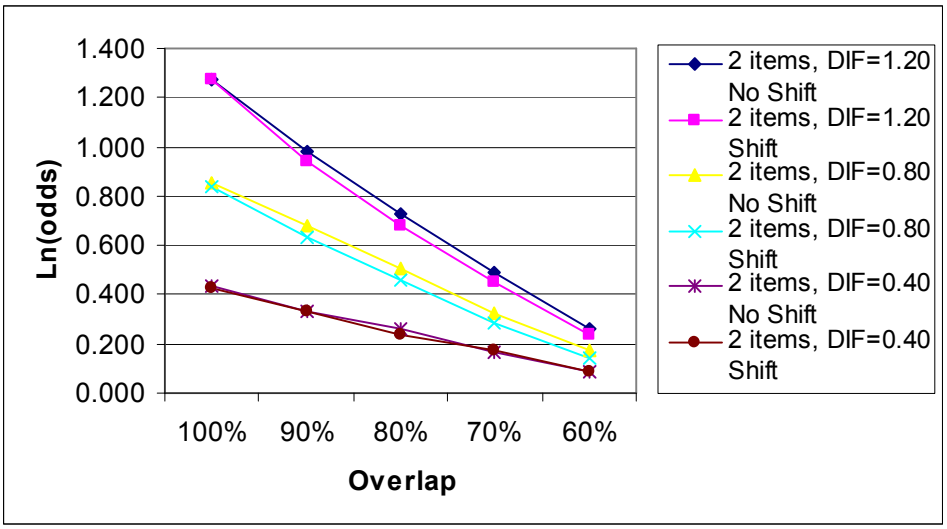
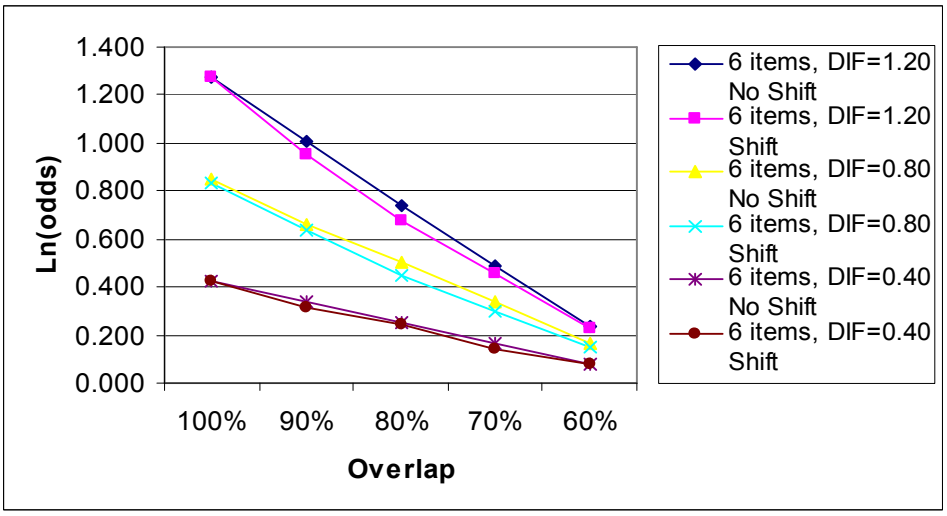


FIGURE 10: Ln(odds) for 50/50 split with 2000 examinees





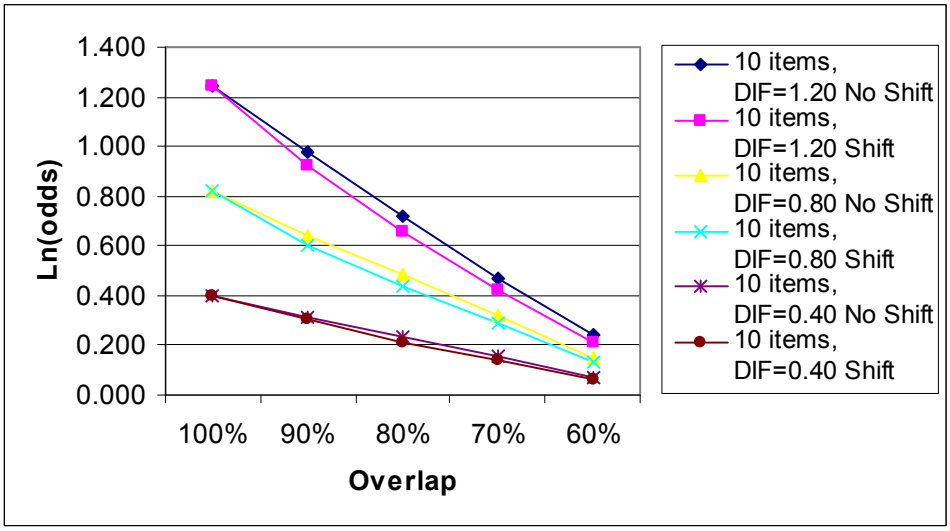


FIGURE 11: Ln(odds) for 80/20 Split with 2000 examinees

Examining Figures 8 through 11 and applying the ETS classification strategy, we see that as the amount of overlap gets smaller it becomes increasingly more difficult to classify an item as having enough DIF to ensure it does not appear on an operational assessment. As shown in Table 4, with overlap less than 80% even items with a large amount of DIF would escape a 'C Classification'.

TABLE 4
The amount of overlap necessary to ensure classification as a B or C (in the ETS classification system) as a function of the magnitude of the differential functioning.

B Classification		C Classification	
Magnitude of DIF	Overlap	Magnitude of DIF	Overlap
1.20	70%	1.20	80%
0.80	80%	0.80	90%
0.40	100%	0.40	Never classified

The average over 100 replications for Ln(odds) is not dependent on sample size or the manifest proportions. The graphs in Figure 12 visually make this point. Likewise, there is no effect due to the contamination of the matching criterion. In each of Figures 8 through 11, moving from the top graph with 10% of the items exhibiting DIF to the bottom with 50% DIF, generally shows the same values for ln(odds). Through examination of these same graphs there does appear to be some effect when the ability distributions of the classes differ. This effect is negligible at best. It should be noted that while the average for ln(odds) is not impacted by the factors listed, the variability of the estimates may be effected. There is more variability when the sample size is smaller, for an 80/20 manifest split rather than a 50/50 split, and for differences in the ability distributions.

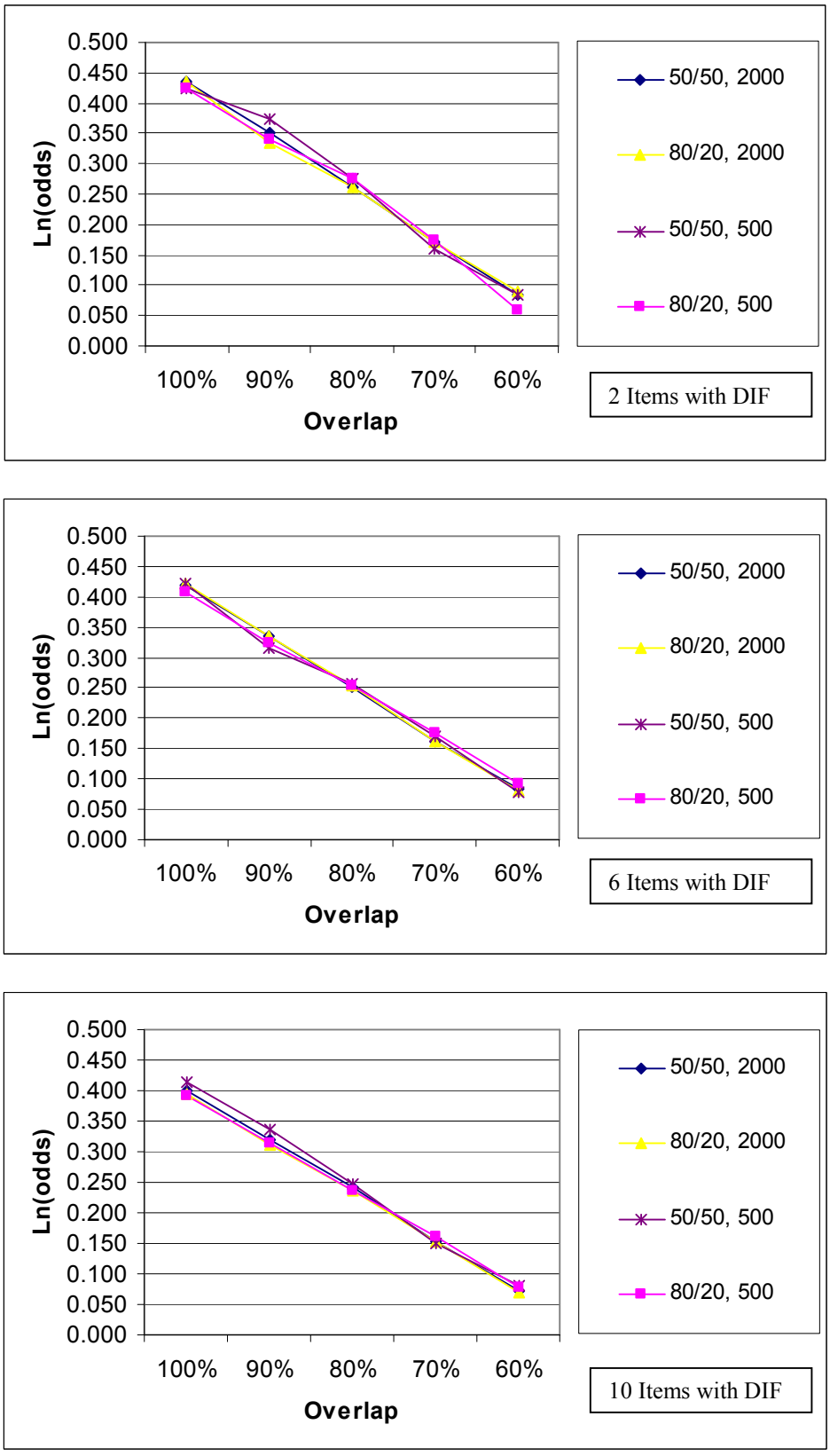


FIGURE 12: Ln(odds) for conditions in which 2, 6 or 10 items contain DIF=0.4

Misclassifications

Results of a regression analysis show that four factors were predictors of the number of items falsely categorized as functioning differentially (i.e. false positives). These were sample size, contamination of the matching criterion as defined by the number items generated to have DIF, degree of overlap, and the manifest proportions. By far the most important factor, accounting for approximately 27% of the variance in misclassifications, was sample size. As had been shown in other studies (Narayanan & Swaminathan, 1994, 1996) more errors occurred as the sample size increased. It should be noted that while an inflation in the number of misclassifications was noted for larger sample size, that rate was never outside of Bradley's (1978) liberal criterion of 0.025 to 0.075 (for alpha = 0.05). Contamination of the matching criterion was positively related to the number of misclassifications. This outcome, including its rather small magnitude, was also found by Narayanan and Swaminathan (1994, 1996) and Fidalgo, et al. (2000). There was also a positive relationship between the degree of overlap between the manifest groups and latent classes and the number of misclassifications. Mean error rates rose steadily from 3.85 per 100 iterations in the 60% overlap condition to 4.26 per 100 iterations when the latent classes and manifest groups were identical. Finally, there was a relationship between manifest proportions and false positives evidenced by the fact that more misclassifications were made in the 50/50 condition than in the 80/20 one.

One surprising result regarding the number of misclassifications of items that were not generated to have DIF was that there did not appear to be a relationship between Type I error rate and the mean difference of the ability distributions. Previous studies (Clauser, et al., 1993; Narayanan & Swaminathan, 1994, 1996) had shown inflation of the

number of misclassifications with unequal ability distributions. This may be due to the manner in which data were generated in those studies. In each case, item difficulty parameters for the focal group were created by adding some positive increment on to the b parameter for the reference group. This seems unrealistic since operational tests generally have some items that advantage the reference group and others that benefit the focal group. For example, on the GRE or SAT (O'Neill & McPeek, 1993) there could be an item with a homograph that would disadvantage minority students along with another question that contains a true cognate advantaging that same group. This method of generating item parameters also exacerbates the differences between the two groups. Starting with mean differences in ability distributions and then having the students of lower ability answer the harder questions will make those students appear even lower in ability. Then, students who should be matched on ability will not be, creating a confound between the ability differences and the effect of the item parameters.

Using an Iterative Procedure

In order to ensure that the findings in this research were not artifacts of the contamination of the matching criterion a small study was undertaken using an iterative Mantel-Haenszel approach. The procedure is consistent with that suggested by Holland and Thayer's (1988) two-stage approach. Step one involves a preliminary DIF analysis to identify potentially bad items. In the second step, a revised total score is calculated that does not include scores on those items, examinees are matched on the purified criterion, and the DIF analysis is repeated. For this analysis, those steps were repeated a second time. This sub-study examined all overlaps for the conditions in which there was a small ($\Delta b=0.40$) amount of DIF on 10 of the 20 items, for 2000 examinees coming from

latent classes with the same ability distributions. Findings (see Figure 13) were consistent with previous research (Zenisky, Hambleton, & Robin, 2003; Fidalgo, et al., 2000). That is, more correct identifications of items functioning differentially were made with the iterative procedure than the standard procedure. In addition, the ln(odds) were the same to two decimal places for the two procedures. The important impression to take away from this sub-study is that the arguments made in the previous section still hold regardless of whether an iterative or a one-step Mantel-Haenszel procedure was used.

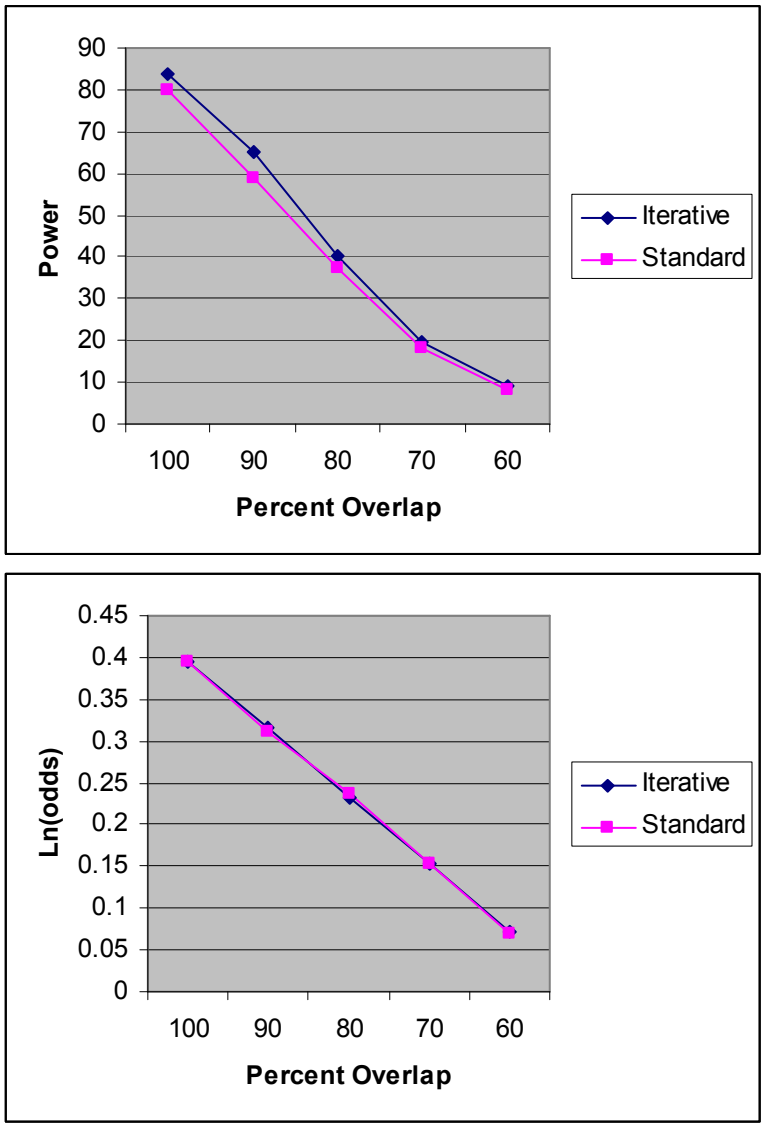


FIGURE 13: Comparisons of output from standard and iterative Mantel-Haenszel procedures

Chapter 5: Results of Simulations with Mixed Rasch Model

Using the mixed Rasch model with the simulated data was important and noteworthy in two ways. First, the results of the BUGS runs illustrated the conditions under which that model could recover the parameters used to generate the data and how well those data were recovered. While this is not a high hurdle to get over, given the model that generated the data matched the model used, if one cannot recover the data there are significant problems. Secondly, using BUGS with the simulated data was useful in terms of clarifying some of the issues that may arise with estimation in these mixture models using MCMC. Each of these is discussed in some depth in the subsequent sections.

Data Recovery

Three aspects of the recovered data can be checked against the gold standard of the simulated data: (1) the items that should and should not function differentially, (2) the estimates of the mean abilities for the two latent classes, and (3) the proportions of males and females within each of the classes. Representative output for the 90% and 60% overlap cases with 500 and 2000 examinees is provided in Tables 5 through 8, and discussed in the subsequent paragraphs relative to those three criteria. Output for those sample sizes for the 80% and 70% conditions, since they provide no further information, are provided in appendix B of this document. In all tables, items correctly identified as functioning differentially (95% confidence interval includes zero) are bolded and false positive identifications are underlined.

TABLE 5
Statistics for 90% overlap on 6 items with different ability distributions (500 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.8687	0.0522	0.001060	0.7560	0.8720	0.9600
Propfemale[2]	0.1313	0.0522	0.001060	0.0400	0.1280	0.2440
Propmale[1]	0.0616	0.0415	8.274E-4	0.0000	0.0560	0.1560
Propmale[2]	0.9384	0.0415	8.274E-4	0.8440	0.9440	1.0000
bdif[1]	-0.6129	0.4429	0.003278	-1.5340	-0.5941	0.2092
bdif[2]	0.2624	0.3248	0.002240	-0.3914	0.2683	0.8839
bdif[3]	**0.7532**	**0.2726**	**0.001832**	**0.2167**	**0.7545**	**1.2830**
bdif[4]	**0.8850**	**0.2388**	**0.001498**	**0.4135**	**0.8855**	**1.3510**
bdif[5]	**0.9866**	**0.2264**	**0.001302**	**0.5435**	**0.9862**	**1.4330**
bdif[6]	0.2049	0.2433	0.001475	-0.2767	0.2059	0.6804
bdif[7]	**-0.7285**	**0.2640**	**0.001776**	**-1.2630**	**-0.7240**	**-0.2240**
bdif[8]	**-0.9378**	**0.2609**	**0.001919**	**-1.4660**	**-0.9318**	**-0.4417**
bdif[9]	**-1.0010**	**0.2666**	**0.001879**	**-1.5390**	**-0.9952**	**-0.4909**
bdif[10]	0.1824	0.2552	0.001546	-0.3199	0.1841	0.6785
bdif[11]	0.1405	0.2791	0.002046	-0.4051	0.1398	0.6893
bdif[12]	-0.0686	0.2525	0.001678	-0.5785	-0.0643	0.4155
bdif[13]	0.1460	0.2349	0.001301	-0.3210	0.1475	0.6058
bdif[14]	0.0698	0.2369	0.001552	-0.3929	0.0691	0.5380
bdif[15]	-0.0777	0.2368	0.001475	-0.5427	-0.0773	0.3819
bdif[16]	0.1805	0.2360	0.001294	-0.2845	0.1808	0.6418
bdif[17]	0.2781	0.2496	0.001397	-0.2152	0.2797	0.7603
bdif[18]	0.0045	0.2890	0.002018	-0.5824	0.0113	0.5510
bdif[19]	-0.2312	0.3059	0.002075	-0.8505	-0.2253	0.3540
bdif[20]	-0.4364	0.3737	0.002728	-1.2270	-0.4152	0.2373
mu[1]	0.0599	0.0912	0.001283	-0.1113	0.0572	0.2459
mu[2]	-1.2300	0.1087	0.001749	-1.4510	-1.2270	-1.0250

Items 3 through 5 and 7 through 9 were generated to have DIF and, as can be seen in Table 5, for the case with 90% overlap, 500 simulated examinees, and differences in ability distributions, those were the only items to function differentially. Recovering the mean abilities for the two latent classes was slightly more problematic. The abilities for

the first latent class were simulated from a standard normal distribution and for the second class from a N(-1,1) distribution. The output from BUGS yields mean abilities of 0.06 and –1.23 respectively. While the true mean for the first class falls within the 95% confidence interval for the posterior mean, this was not the case for the second class. The poor estimates for the second class may be due to the lack of information regarding examinees with low ability because there are no items to differentiate between them. The proportions of males and females in each of the latent classes, though fairly difficult to determine, were adequate estimates of the parameters. Generated to have 90% of the males in one class and 90% of the females in the other, the resultant confidence intervals for the proportions (shown in Table 5) capture the true proportions.

Not surprisingly the estimates for the 60% overlap case with 500 examinees were not as accurate as when the overlap between latent classes and manifest groups was higher. In this situation only three of the six items simulated to have DIF were identified, and all of those items have positive difficulties. See Table 6 for means, standard deviations, MC error and confidence intervals for this case. The means of the ability distributions were estimated to be 0.1458 for the first class, with 60% males, and -1.128 for the second one, with 60% females. The proportions of males in each class were well estimated, while the proportions for the females were less well estimated. This may be due to the fact that females made up a higher percentage of those in the second latent class, which had an ability distribution centered around -1.0. Since there are no items below the item difficulty of -2.0, it is impossible to differentiate examinees at the lowest end of the ability continuum in that class because they will tend to get all of the items incorrect.

TABLE 6
Statistics for 60% overlap on 6 items with different ability distributions (500 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.5538	0.0556	0.001420	0.4600	0.5480	0.6760
Propfemale[2]	0.4462	0.0556	0.001420	0.3240	0.4520	0.5400
Propmale[1]	0.4006	0.0650	0.001892	0.2640	0.4040	0.5120
Propmale[2]	0.5994	0.0650	0.001892	0.4880	0.5960	0.7360
bdif[1]	0.2722	0.4621	0.006730	-0.6973	0.2944	1.1250
bdif[2]	-0.2191	0.4872	0.008584	-1.2320	-0.1981	0.6758
bdif[3]	**0.8737**	**0.3514**	**0.005197**	**0.1608**	**0.8801**	**1.5480**
bdif[4]	**0.9925**	**0.3280**	**0.004215**	**0.3328**	**0.9983**	**1.6170**
bdif[5]	**0.8705**	**0.3157**	**0.003622**	**0.2348**	**0.8784**	**1.4640**
bdif[6]	-0.0607	0.3770	0.006126	-0.8406	-0.0474	0.6412
bdif[7]	-0.7019	0.4436	0.008980	-1.6440	-0.6763	0.1023
bdif[8]	-0.4918	0.3689	0.005495	-1.2570	-0.4775	0.1956
bdif[9]	-0.2056	0.3535	0.005646	-0.9276	-0.1975	0.4648
bdif[10]	-0.2046	0.3721	0.005449	-0.9730	-0.1906	0.4867
bdif[11]	0.2048	0.3764	0.005231	-0.5516	0.2127	0.9203
bdif[12]	-0.7708	0.4305	0.006900	-0.5785	-0.0643	0.4155
bdif[13]	-0.3637	0.3644	0.005453	-1.6680	-0.7515	0.0255
bdif[14]	0.0302	0.3526	0.005528	-0.6905	0.0420	0.6906
bdif[15]	-0.0873	0.2368	0.001475	-0.5427	-0.0706	0.6352
bdif[16]	-0.2765	0.4207	0.008501	-1.1680	-0.2541	0.4863
bdif[17]	-0.0405	0.3872	0.005872	-0.8327	-0.0270	0.6765
bdif[18]	-0.4622	0.4244	0.006061	-1.3530	-0.4432	0.3192
bdif[19]	0.2344	0.4013	0.006735	-0.5914	0.2452	0.9814
bdif[20]	0.4064	0.4925	0.006483	-0.6068	0.4270	1.3090
mu[1]	0.1458	0.1239	0.003070	-0.0939	0.1446	0.3926
mu[2]	-1.1280	0.1256	0.002883	-1.3780	-1.1260	-0.8888

For 2000 examinees rather than 500, the results are predictably better for both the 90% and 60% overlap cases (see Tables 7 and 8). As before, with 90% overlap all six items with DIF were identified, however the standard deviations of the posterior distributions were essentially cut in half. Improvements are noted in the estimates for the

mean of the ability distributions, especially for the latent class with the lower mean ability. However the proportions of males and females in the latent classes show no appreciable improvements in terms of the locations of the posterior distributions.

TABLE 7
Statistics for 90% overlap on 6 items with different ability distributions (2000 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.1354	0.0353	0.001251	0.0700	0.1340	0.2080
Propfemale[2]	0.8646	0.0353	0.001251	0.7920	0.8660	0.9300
Propmale[1]	0.8590	0.0341	0.001182	0.7900	0.8600	0.9240
Propmale[2]	0.1410	0.0341	0.001182	0.0760	0.1400	0.2100
bdif[1]	0.1458	0.2012	0.002460	-0.2425	0.1425	0.5474
bdif[2]	-0.0829	0.1699	0.001846	-0.4137	-0.0846	0.2545
bdif[3]	**-0.7431**	**0.1400**	**0.001507**	**-1.0170**	**-0.7429**	**-0.4696**
bdif[4]	**-0.9359**	**0.1317**	**0.001453**	**-1.1950**	**-0.9350**	**-0.6791**
bdif[5]	**-0.7594**	**0.1244**	**0.001260**	**-1.0040**	**-0.7589**	**-0.5169**
bdif[6]	-0.0872	0.1284	0.001240	-0.3393	-0.0870	0.1643
bdif[7]	**1.2140**	**0.1557**	**0.002201**	**0.9188**	**1.2090**	**1.5320**
bdif[8]	**0.7334**	**0.1325**	**0.001339**	**0.4765**	**0.7324**	**0.9953**
bdif[9]	**1.0910**	**0.1471**	**0.001817**	**0.8079**	**1.0900**	**1.3830**
bdif[10]	-0.0157	0.1359	0.001343	-0.2870	-0.0165	0.2543
bdif[11]	0.0075	0.1480	0.001576	-0.2795	0.0068	0.3002
bdif[12]	0.0027	0.1380	0.001354	-0.2664	0.0024	0.2764
bdif[13]	0.0109	0.1283	0.001150	-0.2407	0.0105	0.2638
bdif[14]	0.1803	0.1276	0.001277	-0.0685	0.1799	0.4314
bdif[15]	-0.1753	0.1248	0.001169	-0.4194	-0.1760	0.0670
bdif[16]	-0.2126	0.1248	0.001216	-0.4570	-0.2133	0.0334
bdif[17]	0.0353	0.1353	0.001336	-0.2273	0.0342	0.3019
bdif[18]	-0.2099	0.1402	0.001413	-0.4831	-0.2106	0.0670
bdif[19]	0.0244	0.1558	0.001775	-0.2767	0.0233	0.3356
bdif[20]	-0.2231	0.1798	0.001415	-0.5733	-0.2252	0.1331
mu[1]	-1.0430	0.0567	0.001638	-1.1590	-1.0410	-0.9362
mu[2]	0.0456	0.0542	0.001438	-0.0579	0.0447	0.1556

For the case with 60% overlap with 2000 examinees, the six items with DIF were identified, the mean abilities well estimated, and the proportions of males within the classes were better estimated than those of the females. The lone difference between this condition and the 90% overlap condition was the misclassification of one item (#10).

TABLE 8
Statistics for 60% overlap on 6 items with different ability distributions (2000 examinees)

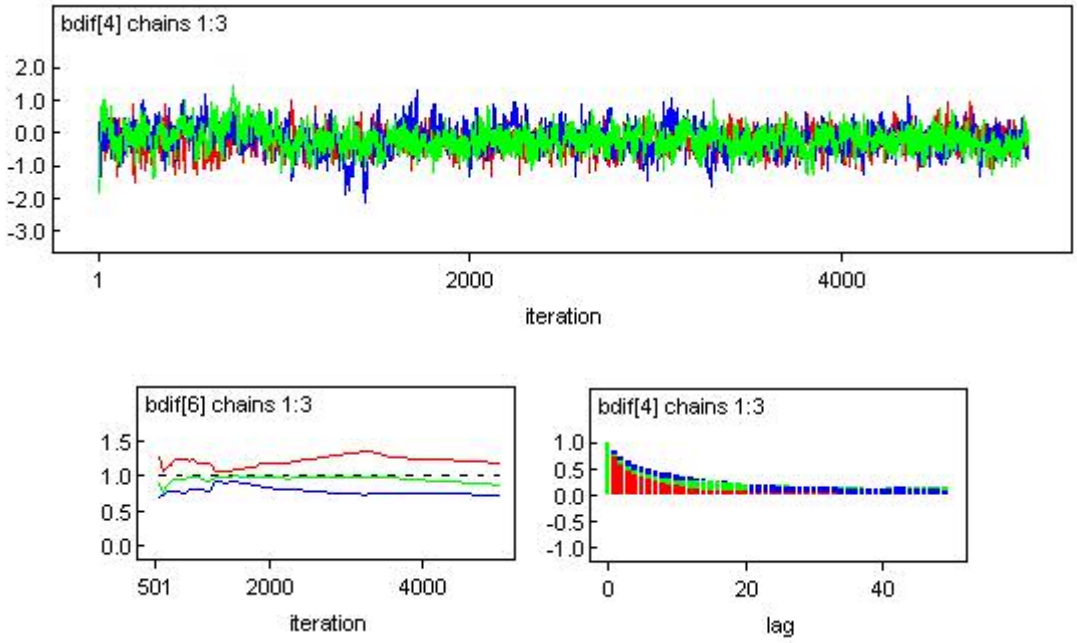
node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.3860	0.0617	0.002609	0.2700	0.3850	0.4950
Propfemale[2]	0.6140	0.0617	0.002609	0.5050	0.6150	0.7300
Propmale[1]	0.5944	0.0595	0.002516	0.4970	0.5900	0.7100
Propmale[2]	0.4056	0.0595	0.002516	0.2900	0.4100	0.5030
bdif[1]	-0.1068	0.2666	0.004916	-0.6188	-0.1112	0.4249
bdif[2]	0.0977	0.2596	0.004861	-0.3885	0.0899	0.6327
bdif[3]	**-0.7528**	**0.1860**	**0.002716**	**-1.1140**	**-0.7533**	**-0.3848**
bdif[4]	**-0.9529**	**0.1694**	**0.002483**	**-1.2830**	**-0.9526**	**-0.6198**
bdif[5]	**-0.8477**	**0.1663**	**0.002494**	**-1.1710**	**-0.8493**	**-0.5178**
bdif[6]	0.0658	0.1840	0.002928	-0.2835	0.0621	0.4401
bdif[7]	**0.5119**	**0.2170**	**0.004197**	**0.1004**	**0.5060**	**0.9485**
bdif[8]	**0.8532**	**0.2484**	**0.005617**	**0.3852**	**0.8460**	**1.3560**
bdif[9]	**0.8306**	**0.2288**	**0.004567**	**0.3968**	**0.8262**	**1.2910**
bdif[10]	-0.4427	0.1971	0.003798	-0.8230	-0.4459	-0.0502
bdif[11]	0.0145	0.2134	0.003839	-0.3927	0.0082	0.4452
bdif[12]	-0.2942	0.1942	0.003289	-0.6664	0.2979	0.0931
bdif[13]	0.0308	0.2002	0.003577	-0.3465	0.0261	0.4384
bdif[14]	0.3462	0.1991	0.003675	-0.0345	0.3431	0.7464
bdif[15]	-0.1126	0.1829	0.003005	-0.4666	-0.1147	0.2517
bdif[16]	-0.1088	0.1815	0.002994	-0.4600	-0.1113	0.2564
bdif[17]	0.3772	0.2174	0.004149	-0.0360	0.3736	0.8121
bdif[18]	-0.1696	0.1938	0.003092	-0.5429	-0.1726	0.2206
bdif[19]	0.2849	0.2443	0.005112	-0.1678	0.2760	0.7932
bdif[20]	0.3752	0.3164	0.006529	-0.1843	0.3528	1.0650
mu[1]	-1.0150	0.0831	0.003128	-1.1860	-1.0100	-0.8657
mu[2]	-0.0109	0.0799	0.002870	-0.1574	-0.0127	0.1501

Issues in Using MCMC for these Models

In addition to clarifying the conditions under which the mixed Rasch model was effective in detecting DIF, issues arising when estimating mixture models using MCMC were uncovered. For the mixed Rasch model employed in this research, it appeared that some parameters in the model were much better determined than others. This is discussed in some depth in the subsequent paragraphs. Additionally, dependencies within chains for some parameters were discovered. The impact these had on the number of iterations necessary for burn-in and the strategy used for dealing with this problem are also discussed below.

Not surprisingly the Rasch item difficulty differences were the easiest to estimate of the parameters under consideration. Examinations of time series plots indicated that the item difficulty differences (bdif) typically converged within 1,000 iterations for all simulated conditions. However, inspection of the BGR diagnostic plots for bdif provided evidence that more iterations were necessary to achieve convergence when sample size was small. For the conditions with 500 examinees, 5,000 iterations were generally necessary, and for 2,000 examinees typically only 1,000 iterations were needed. Examination of the autocorrelation plots for the difficulty parameters indicated that for estimation of these parameters it would not be necessary to run an extremely large number of iterations to traverse the entire sample space. Figure 14 shows representative history plots, BGR diagnostic plots, and graphs of autocorrelations for the bdif parameters.

For simulated data with 500 examinees



For simulated data with 2,000 examinees

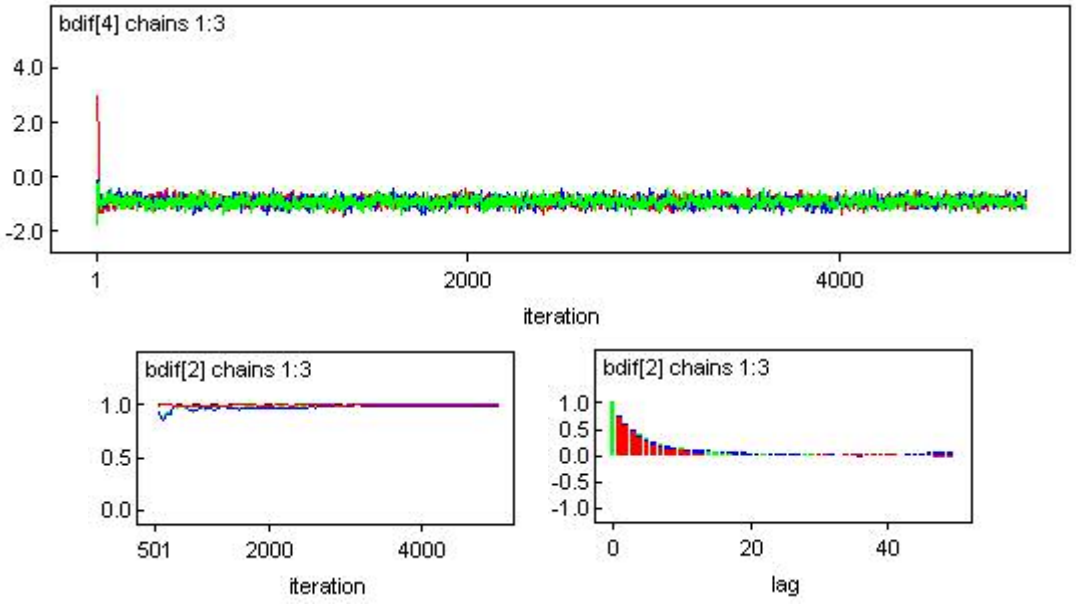


FIGURE 14: Diagnostic plots for bdif

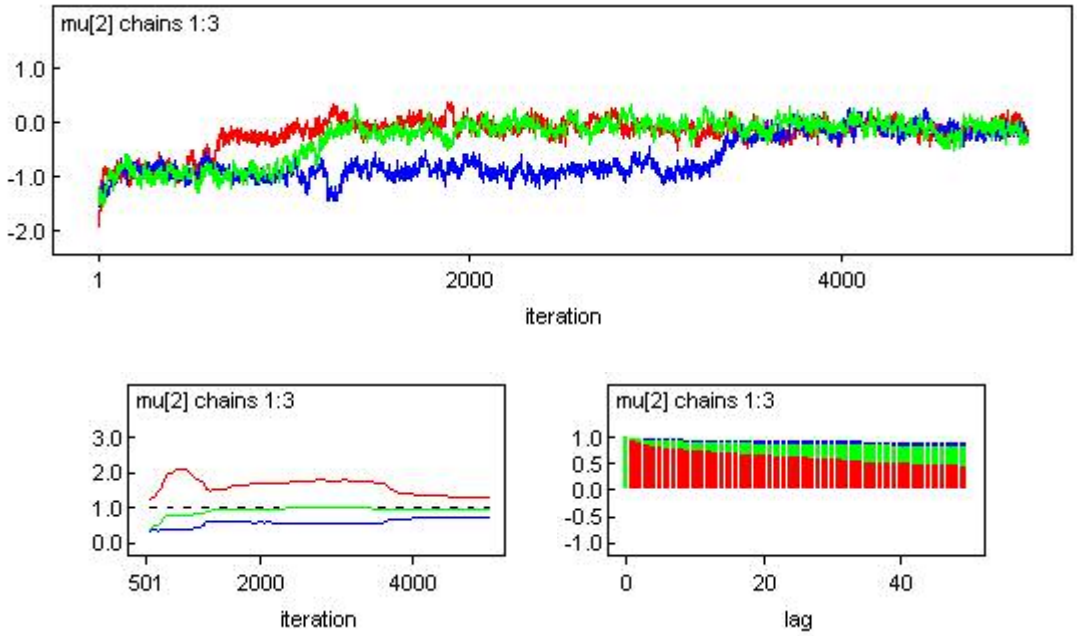
The means of the latent ability distributions proved to be more problematic to estimate than the differences in the item difficulties. Based on the time series plots it appeared that these means for the conditions with 500 examinees required a burn-in of approximately 4,000 iterations, but the BGR diagnostic plots again showed that more iterations were necessary. Typically 5,000 iterations were sufficient to get all curves to go to one. For the conditions with 2,000 examinees the time series plots stabilized within a few hundred iterations and the BGR diagnostics showed stationarity had been reached by 2,500 iterations. The one noteworthy indicator under both sample sizes was that the autocorrelation remained high, probably due to the cross-correlation among the latent ability and other parameters in the model. See Figure 15 for typical diagnostic plots for the means of the latent ability distributions.

The most challenging parameters to estimate were the proportion of males and females within each latent class. Convergence of the chains based on the BGR diagnostic plots typically took place for the parameters within 2,500 (n = 2000) to 5,000 (n = 500) iterations. As is shown in Figure 16, the time series plots indicated that while there seems to be convergence within several hundred iterations for the larger sample size, 4,000 iterations were required when there were fewer examinees. However, while these chains did seem to converge around the true probabilities, it was clear from the high autocorrelations and the "wandering" nature of the time series plots that the draws were not random (see Figure 17 for a comparison of chains that wander and those that do not). Though the strategy of thinning, or discarding some iterations, may once have been considered, Gilks, Richardson and Spiegelhalter (1996, pg. 140) now say "there is no advantage in discarding intermediate simulation draws, even if highly correlated."

Instead the state of the art seems to be to pool draws from a number of chains. The fact that the BGR diagnostic had stabilized around one, indicating that the means of the chains were essentially the same, lends credence to the idea of pooling results of independent chains in this case. Additionally, compressing the scale for a "wandering" chain and letting it run longer would yield a graph that would provide more evidence that the chain had converged, albeit with a higher autocorrelation.

The lessons from these simulations are the following. First, a relatively large number of iterations were necessary as burn-in. For sample sizes similar to those used in this study a burn-in of 5,000 iterations should be sufficient for a 20 item test. For tests with more items a shorter burn-in might be appropriate but the same issues regarding dependence between some chains would remain. Second, it is vital to run multiple chains. Besides the fact that it is more likely to sample the entire posterior when several chains with divergent starting values are used, there should be no dependence between chains. Since there is dependence within chains, pooling the draws from independent chains should ameliorate the impact of that dependence to some degree. Finally it seems prudent to end up with a sample of approximately 50,000 in order to feel comfortable making inferences regarding the posterior distributions. When that was done the density plots were smooth (see Figure 2 on page 24) and the standard deviation to MC-error ratio was less than the recommended ratio of 0.05 (Spiegelhalter, et al., 2003).

For simulated data with 500 examinees



For simulated data with 2,000 examinees

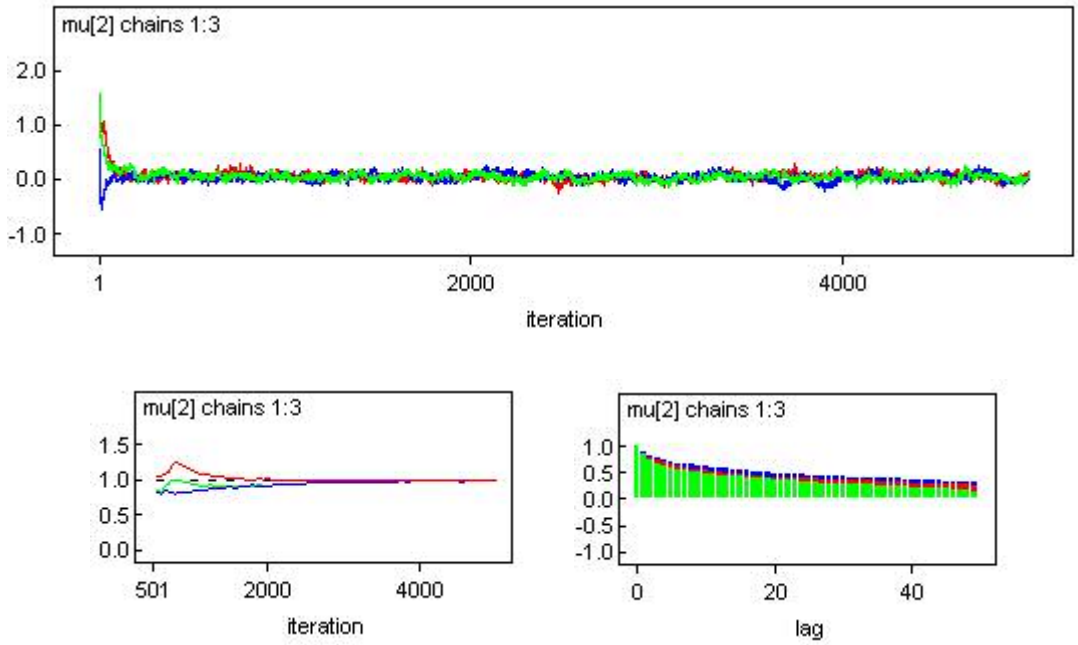
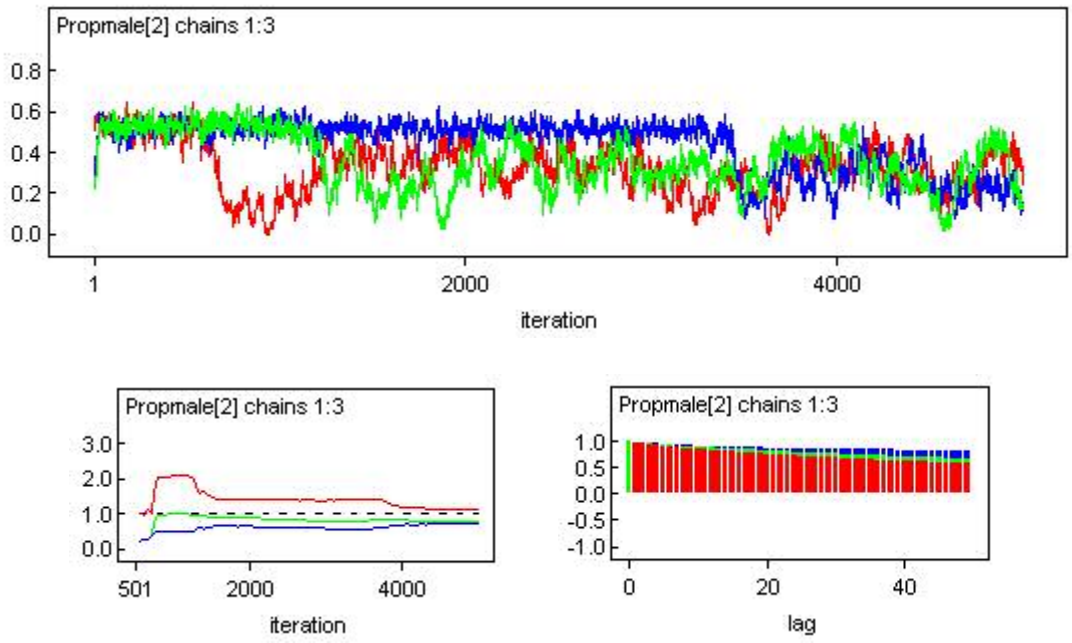


FIGURE 15: Diagnostic plots for the latent ability distributions

For simulated data with 500 examinees



For simulated data with 2,000 examinees

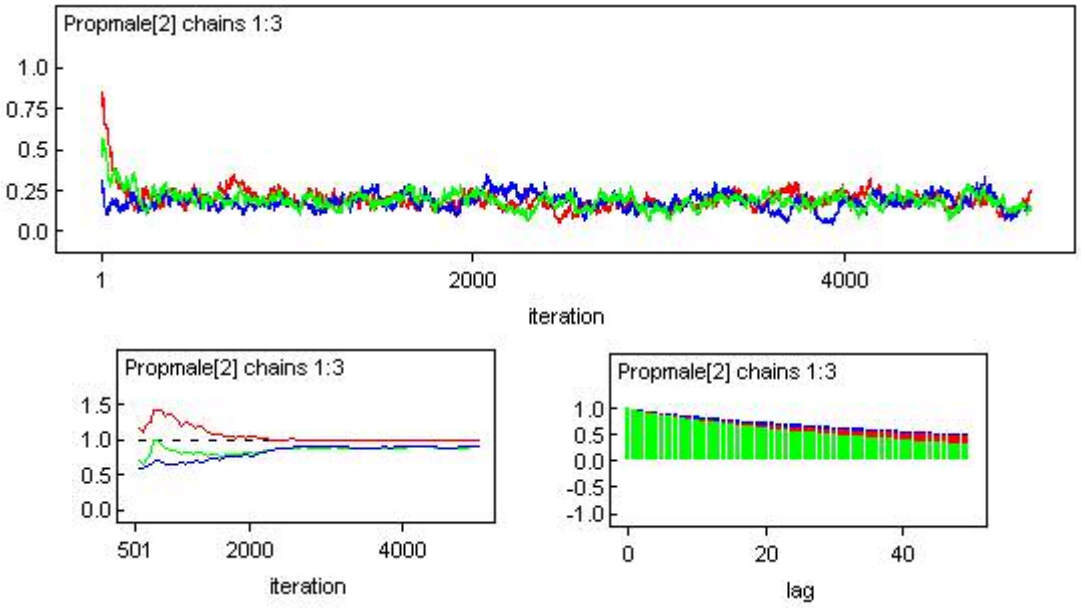


FIGURE 16: Diagnostic plots for proportions of manifest groups within latent classes

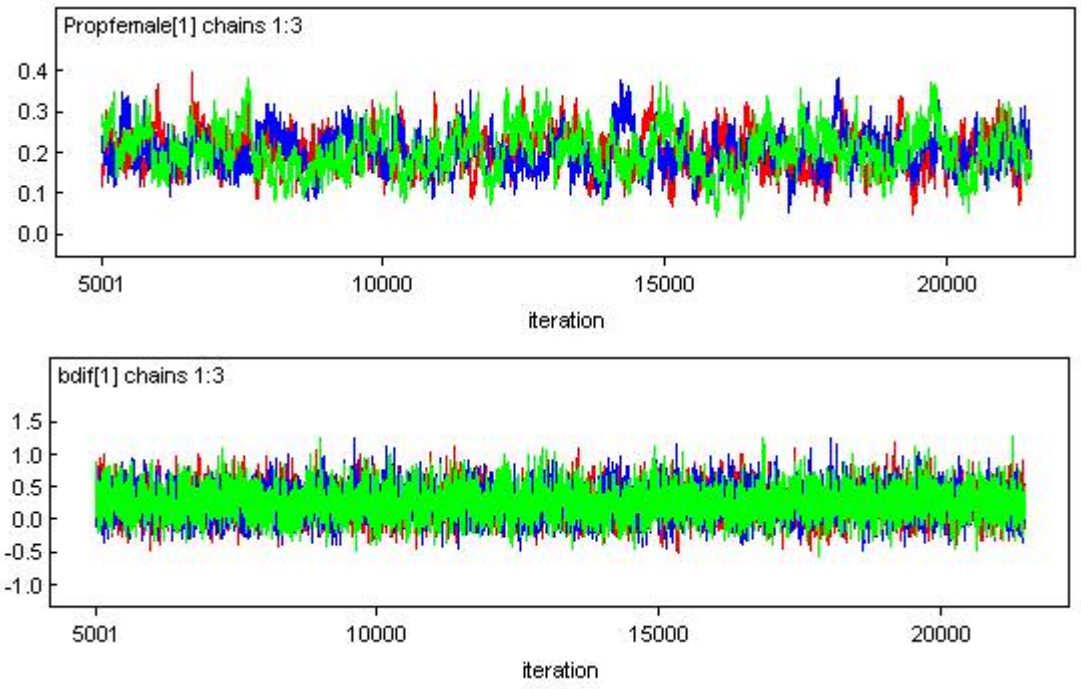


FIGURE 17: Comparison of "wandering" time series plot (top) and one that indicates random sampling (bottom) within the same part of the same space for all chains

Chapter 6: Results of Data Analysis

Background information

The data used in this study were a subset of responses for 1016 students on Form A of an English language proficiency test in reading for grade cohort 3-5. Though this form originally contained 60 items, 15 of those items were judged to be suspect by the vendor after the piloting of the test. Of the remaining 45 items, the questions pertaining to the final two long reading selections were dropped since there was evidence that many students did not reach those items on the test. Therefore, only 34 items were retained for this analysis.

Frequencies for this sample for the categorical variables collected in this study are shown in the table below. The Asian students tested represented a variety of countries including China, Japan, Vietnam and Korea. Hispanic students came from countries in Central and South America, the Caribbean and Europe. It is also interesting to note that a large number of these students were English language learners born in the United States.

TABLE 9
Frequencies for categorical variables

		Frequency	Percent
Gender	Male	479	47.1
	Female	537	52.9
Grade	Third	355	34.9
	Fourth	367	36.1
	Fifth	294	28.9
Ethnicity	Asian	136	13.4
	Hispanic	880	86.6
Born in US	No	585	57.6
	Yes	431	42.4

These data were analyzed using the Mantel-Haenszel technique to uncover the items that would be identified as functioning differentially based on the current DIF techniques. Four items, numbers 18, 25, 30, and 34, had statistically significant DIF with items 18 and 30 advantaging the Asian students and the other two items advantaging the Hispanic examinees. Six items, numbers 7, 9, 23, 27, 33, and 34 showed statistically significant gender DIF; the first three of these items favored females and the last three advantaged male examinees. See Table 10 on the next page for the results of these analyses.

Protocols and Results

A four step approach was defined for examining differential item functioning using a latent class perspective. Those steps are:

1. Identify the model that best fits the data;
2. Decide whether the manifest group percentages within the latent classes warrant a latent class approach;
3. Examine the data from the latent class analysis for clues as to why there is DIF and to inform the choice of covariates;
4. Use the covariates to predict membership in the latent classes.

Each of these steps will be discussed below, along with the results of that step from the analyses of these data.

An important first step in the analysis of items for DIF using the approach under consideration is to assess which latent class model or models fit the data. For each model one asks, is this a reasonable approximation to the observed data? Models that do not fit the data may be set aside and the choice of the most appropriate model can be made from those that do appear to fit.

TABLE 10
Results of Mantel-Haenszel analyses

	Gender			Ethnicity		
Item	**MH-χ^2**	**Sig.**	**Ln(odds)**	**MH-χ^2**	**Sig.**	**Ln(odds)**
bdif[1]	3.621	0.057	-0.490	1.058	0.304	0.761
bdif[2]	0.981	0.322	-0.361	2.668	0.102	2.031
bdif[3]	0.000	0.995	-0.051	0.063	0.802	0.470
bdif[4]	0.647	0.421	-0.168	0.002	0.960	-0.057
bdif[5]	1.960	0.162	-0.270	0.329	0.566	-0.207
bdif[6]	2.714	0.099	-0.273	0.887	0.346	-0.233
bdif[7]	**6.946**	**0.008**	**0.697**	0.077	0.782	-0.238
bdif[8]	0.928	0.335	0.227	1.091	0.296	-0.406
bdif[9]	**5.407**	**0.020**	**0.452**	0.263	0.608	-0.195
bdif[10]	1.444	0.229	0.337	0.321	0.571	0.355
bdif[11]	0.052	0.820	0.066	0.002	0.966	0.048
bdif[12]	0.415	0.519	0.152	1.353	0.245	0.510
bdif[13]	0.049	0.825	0.045	1.233	0.267	0.304
bdif[14]	1.871	0.171	0.233	0.029	0.865	0.077
bdif[15]	0.065	0.799	-0.061	0.172	0.679	0.166
bdif[16]	0.732	0.392	0.163	1.135	0.287	-0.325
bdif[17]	0.749	0.387	0.142	0.069	0.792	0.093
bdif[18]	0.010	0.921	0.000	**4.322**	**0.038**	**0.879**
bdif[19]	3.438	0.064	0.500	0.089	0.765	0.244
bdif[20]	1.094	0.295	-0.199	0.020	0.886	0.083
bdif[21]	0.074	0.786	-0.068	0.002	0.987	0.043
bdif[22]	0.642	0.423	0.209	0.005	0.946	0.113
bdif[23]	**5.138**	**0.023**	**0.569**	0.002	0.969	-0.077
bdif[24]	0.392	0.531	0.114	0.431	0.512	-0.202
bdif[25]	0.040	0.841	0.055	**8.096**	**0.004**	**-0.995**
bdif[26]	0.683	0.409	0.187	0.005	0.944	-0.095
bdif[27]	**15.985**	**0.000**	**-0.643**	0.011	0.917	0.048
bdif[28]	1.214	0.270	0.200	0.562	0.453	-0.230
bdif[29]	2.285	0.131	-0.260	0.970	0.325	0.301
bdif[30]	0.046	0.831	0.045	**4.807**	**0.028**	**0.539**
bdif[31]	1.190	0.275	0.186	0.001	0.970	-0.023
bdif[32]	2.871	0.090	-0.281	0.705	0.401	-0.233
bdif[33]	**4.717**	**0.030**	**-0.330**	0.326	0.568	0.145
bdif[34]	**3.997**	**0.046**	**-0.321**	**8.743**	**0.003**	**-0.702**

In this case, since the data examined had item responses for males and females who were Asian or Hispanic, it was reasonable to test one-, two-, and three-class models. In addition to statistically checking for fit, it was appropriate to be concerned with the meaningfulness of the latent classes. This could be done by considering the percentage of examinees within a class or based upon some substantive rationale. Since a substantive rationale did not exist in this case, percentages were used to judge whether or not a class was worth retaining. All other considerations being equal, the most parsimonious model was chosen.

The results of the three model-fit analyses done using the shadow data technique discussed in the chapter 3 are shown below. Evidence was provided that the one-class model did not fit the data since the 95% confidence interval for the proportion of times the observed data was worse than the shadow data spans from 0.5266 to 0.5827. Since this interval did not include 0.5, the proportion one would expect by chance, there was evidence that the one-class model did not fit the data. This result was expected given the complexity of the data in terms of the genders and ethnic backgrounds of the examinees. There is, however, evidence that both the two- and three-class models fit the data since both confidence intervals include 0.5. Examination of the numbers of students for the 3-class model showed less than 10 Asian students in one of the classes. Given that small number, the more parsimonious model was chosen.

TABLE 11
Model fit for 1-, 2- and 3-class models using the shadow data technique

	mean	sd	MC error	2.5%	median	97.5%	start	sample
3 class	0.4893	0.0154	1.099E-4	0.4594	0.4895	0.5196	501	50000
2 class	0.5159	0.0152	8.250E-5	0.4865	0.5155	0.5456	501	50000
1 class	0.5551	0.1439	7.923E-5	0.5266	0.5547	0.5827	501	50000

The next step in the procedure is really a decision node. Once the percentage of people in each manifest group in the latent classes had been found, a determination needed to be made regarding the appropriateness of using the manifest group as a proxy for the latent group. Most would agree that if there was 99% overlap between latent and manifest groups it would be appropriate to say the item functioned differentially against that manifest group. Conversely, most people would say it would be inappropriate to use the manifest groups if there was only 60% overlap, meaning 40% of the people in the manifest groups behaved like those in the other manifest group. Though this research will not impose a decision about the cut-off for when that overlap is large enough, it is important to note that this decision must be made.

The first latent class was made up of 90.8% of the Asian females, and 74.9% of the Asian males, yielding approximately 83% of the Asian students in that latent class. That latent class also included 82.0% of the Hispanic females, and 64.9% of the Hispanic males, meaning that 74% of the Hispanic students were in the first class. Additionally we note that the first class consisted of 83% of the females and 66% of the male examinees. Overall, approximately three quarters of the examinees were in the first latent class and the remaining one-quarter in the second class.

The percentages regarding the ethnicity and sex of the students within the classes provided evidence that a latent class perspective was called for since they show that ethnicity and sex were not important indicators of why student response patterns differed on this test. Further proof of this came when ethnicity, sex, and the interaction of the two were used to predict membership in the two latent classes, and none of the regression coefficients were significant. It is interesting to note that an examination of the small

differences between the percentages based on ethnicity and sex showed that the Asian and Hispanic students responded more similarly than males and females. This may explain why the Mantel-Haenszel procedure identified more items functioning differentially between males and females than between Asians and Hispanics.

The third step in this procedure was to examine the data from the latent class analysis for clues as to why there were items that functioned differentially. This was done by analyzing the following:

1. Mean abilities within the latent class;
2. Magnitudes of the differences in item difficulties between the latent classes;
3. Patterns of item difficulties within classes.

One would expect the results of these analyses, together with an understanding of both the examinees and the test, to yield some clues as to why there was DIF.

Mean Abilities

Table 12 on the next page shows the mean abilities and standard deviations for the manifest groups within the latent classes. Looking at this data at the level of the latent classes we see that examinees in the first latent class tend to be much more able than those in the second class. Additionally, it seems that the Asian examinees in the first class (mean ability of 2.7645) were, on average, significantly more able than the Hispanic students in that class (mean ability of 2.2875). Though the same patterns appears to hold for Asian and Hispanic examinees in the second class, the small number of Asian students in this class yielded estimates that were too unstable from which to make generalizations. There were no significant differences between the mean abilities of the males and females in the two classes.

TABLE 12
Mean abilities (standard deviations) for manifest groups within latent classes

	Latent Class #1		Latent Class #2	
	Female	Male	Female	Male
Asian	2.853 (0.1514)	2.662 (0.2113)	-0.2547 (0.5398)	+0.1385 (0.3304)
Hispanic	2.322 (0.0797)	2.247 (0.1082)	-0.6589 (0.1608)	-0.1165 (0.1319)

In this case, the mean ability estimates provided little assistance with regard to identifying the nature of the latent classes. Though small differences existed between Asian and Hispanic examinees, a general trend holds – that examinees in the first class were, on average, much more able than those in the second class. From this we know that if latent class membership was predicated on strategy usage (as an example), the strategy used by the students in the second latent class was much less effective than the one used by the students in the first class.

Magnitudes of Differences in Item Difficulties

Table 13 shows the items identified as functioning differentially from a latent class perspective, including the magnitude of the differential item functioning (shown as the mean for bdif) for each of the 34 items on this test. Items that are bolded are those for which the confidence interval for the difference between the item difficulties in the two classes does not contain zero. In contrast to the DIF results from the Mantel-Haenszel procedure, the majority of the items function differently for the two latent classes, with 23 of the 34 items exhibiting statistically significant DIF. It should be noted, however, that some of these items, like question #14, may not have meaningful amounts of DIF (i.e. the ln(odds) may be relatively low).

TABLE 13
Item difficulties from the latent class analysis

Node	Mean	Stat. DIF	MC Error	2.5%	Median	97.5%
bdif[1]	0.0290	0.2631	0.004307	-0.4994	0.0331	0.5361
bdif[2]	**-1.3840**	**0.4488**	**0.010890**	**-2.3130**	**-1.3700**	**-0.5468**
bdif[3]	**-1.6650**	**0.4922**	**0.013880**	**-2.6920**	**-1.6394**	**-0.7723**
bdif[4]	**0.9019**	**0.2184**	**0.003205**	**0.4705**	**0.9034**	**1.3270**
bdif[5]	0.1703	0.2171	0.002725	-0.2580	0.1723	0.5914
bdif[6]	**0.6153**	**0.2036**	**0.002497**	**0.2137**	**0.6174**	**1.0100**
bdif[7]	**-1.3670**	**0.3387**	**0.006702**	**-2.0628**	**-1.3570**	**-0.7337**
bdif[8]	**-1.1160**	**0.2741**	**0.004241**	**-1.6770**	**-1.1100**	**-0.5999**
bdif[9]	-0.0798	0.2267	0.003026	-0.5317	-0.0774	0.3585
bdif[10]	**-0.9592**	**0.3145**	**0.006053**	**-1.6000**	**-0.9490**	**-0.3642**
bdif[11]	**-0.8664**	**0.2622**	**0.004845**	**-1.4010**	**-0.8587**	**-0.36791**
bdif[12]	**-1.1170**	**0.2714**	**0.004955**	**-1.6680**	**-1.1090**	**-0.6028**
bdif[13]	**1.0700**	**0.1978**	**0.002438**	**0.6771**	**1.0730**	**1.4520**
bdif[14]	**0.5254**	**0.2075**	**0.002408**	**0.1130**	**0.5275**	**0.9268**
bdif[15]	-0.1163	0.2217	0.002974	-0.5530	-0.1149	0.3161
bdif[16]	-0.2372	0.2203	0.002843	-0.6757	-0.2345	0.1859
bdif[17]	**0.9426**	**0.2032**	**0.003112**	**0.5378**	**0.9446**	**1.3350**
bdif[18]	**-0.6141**	**0.2416**	**0.003483**	**-1.0990**	**-0.6119**	**-0.1455**
bdif[19]	**-0.9514**	**0.3278**	**0.006156**	**-1.6240**	**-0.9424**	**-0.3351**
bdif[20]	0.1718	0.2204	0.003707	-0.2700	0.1736	0.6000
bdif[21]	-0.2858	0.2347	0.004010	-0.7510	-0.2841	0.1704
bdif[22]	**-0.7487**	**0.2814**	**0.005254**	**-1.3090**	**-0.7444**	**-0.2096**
bdif[23]	**-0.8467**	**0.3127**	**0.005524**	**-1.4840**	**-0.8362**	**-0.2620**
bdif[24]	0.4488	0.2305	0.003704	-0.0064	0.4508	0.8916
bdif[25]	0.0132	0.2357	0.005048	-0.4598	0.0141	0.4675
bdif[26]	-0.3768	0.2263	0.003429	-0.8264	-0.3750	0.0608
bdif[27]	**0.9935**	**0.2051**	**0.003020**	**0.5891**	**0.9936**	**1.3940**
bdif[28]	0.1673	0.2254	0.003280	-0.2815	0.1698	0.6053
bdif[29]	**09989**	**0.1983**	**0.002503**	**0.6108**	**0.9985**	**1.3890**
bdif[30]	**1.2530**	**0.2151**	**0.002462**	**0.8264**	**1.2560**	**1.6670**
bdif[31]	**0.7181**	**0.2071**	**0.003001**	**0.3079**	**0.7185**	**1.1230**
bdif[32]	**0.7094**	**0.2008**	**0.002212**	**0.3091**	**0.7106**	**1.0990**
bdif[33]	**1.7870**	**0.1999**	**0.002614**	**1.3960**	**1.7870**	**2.1770**
bdif[34]	**1.2150**	**0.2044**	**0.002550**	**0.8105**	**1.2160**	**1.6130**

At this point, comparing the differences in the item difficulties to the characteristics of the items may provide information regarding the identity of the latent classes. Content matter experts could categorize the items based on differences such as the level of cognitive thinking or the type of prompt involved. For this data set, the items were classified using Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) and the questions were categorized using general distinctions provided by the test publisher and more specific features regarding the stimulus material or response options. These general and specific types of items are:

- Short passages – with either a picture or short paragraph as the prompt
- Instructions – response options may be either graphic or verbal choices
- Longer passages – with either a schedule or a long passage as the prompt

The categorizations of the items are shown in the table on the next page. Note that these classifications are not meant to be an exhaustive list of categories but rather a representative sample. Content and cognitive experts would undoubtedly be able to provide many more meaningful ways to categorize these items.

With respect to the level of cognitive thinking and the general prompt type there are no clear patterns. For example, items at the knowledge level of Bloom's taxonomy exhibit latent DIF in some cases but not in others. However, one key to the makeup of the latent classes does come to light through this sort of analysis of item features. It is clear that items 29 through 34, which all show latent DIF favoring examinees in the first latent class, refer to one reading passage on the test. On its own this piece of information may not provide enough evidence to determine why two latent classes exist, but this may prove valuable in concert with other pieces of information.

TABLE 14
Item characteristics and latent DIF

Item Number	Difference in Item Difficulties	Level of Cognition	General Type of Prompt	Specific Type of Prompt
1	0.0290	Knowledge	Short	Picture
2	**-1.3840**	**Knowledge**	**Short**	**Picture**
3	**-1.6650**	**Knowledge**	**Short**	**Picture**
4	**0.9019**	**Comprehension**	**Short**	**Picture**
5	0.1703	Comprehension	Short	Picture
6	**0.6153**	**Comprehension**	**Short**	**Picture**
7	**-1.3670**	**Comprehension**	**Short**	**Paragraph**
8	**-1.1160**	**Knowledge**	**Short**	**Paragraph**
9	-0.0798	Knowledge	Short	Paragraph
10	**-0.9592**	**Comprehension**	**Short**	**Paragraph**
11	**-0.8664**	**Knowledge**	**Short**	**Paragraph**
12	**-1.1170**	**Knowledge**	**Short**	**Paragraph**
13	**1.0700**	**Comprehension**	**Instructions**	**Graphic Response**
14	**0.5254**	**Comprehension**	**Instructions**	**Graphic Response**
15	-0.1163	Comprehension	Instructions	Graphic Response
16	-0.2372	Comprehension	Instructions	Written Response
17	**0.9426**	**Comprehension**	**Instructions**	**Written Response**
18	**-0.6141**	**Comprehension**	**Long**	**Schedule**
19	**-0.9514**	**Comprehension**	**Long**	**Schedule**
20	0.1718	Application	Long	Schedule
21	-0.2858	Comprehension	Long	Schedule
22	**-0.7487**	**Comprehension**	**Long**	**Schedule**
23	**-0.8467**	**Comprehension**	**Long**	**Schedule**
24	0.4488	Comprehension	Long	Passage (1)
25	0.0132	Comprehension	Long	Passage (1)
26	-0.3768	Analysis	Long	Passage (1)
27	**0.9935**	**Knowledge**	**Long**	**Passage (1)**
28	0.1673	Comprehension	Long	Passage (1)
29	**0.9989**	**Comprehension**	**Long**	**Passage (2)**
30	**1.2530**	**Knowledge**	**Long**	**Passage (2)**
31	**0.7181**	**Knowledge**	**Long**	**Passage (2)**
32	**0.7094**	**Comprehension**	**Long**	**Passage (2)**
33	**1.7870**	**Comprehension**	**Long**	**Passage (2)**
34	**1.2150**	**Analysis**	**Long**	**Passage (2)**

Patterns of Item Difficulties

Examination of Figure 18 shows a clear pattern of item difficulties; for the first latent class the first third of the items tend to be easier and last third harder than for the second class. One could hypothesize the reason for this is that the initial items on the test deal with shorter amounts of concrete information while the final items ask students to consider longer, more complex stimulus materials. Therefore students who tend to memorize factual knowledge will do very well on the first group of items; however they will perform poorly on the later items.

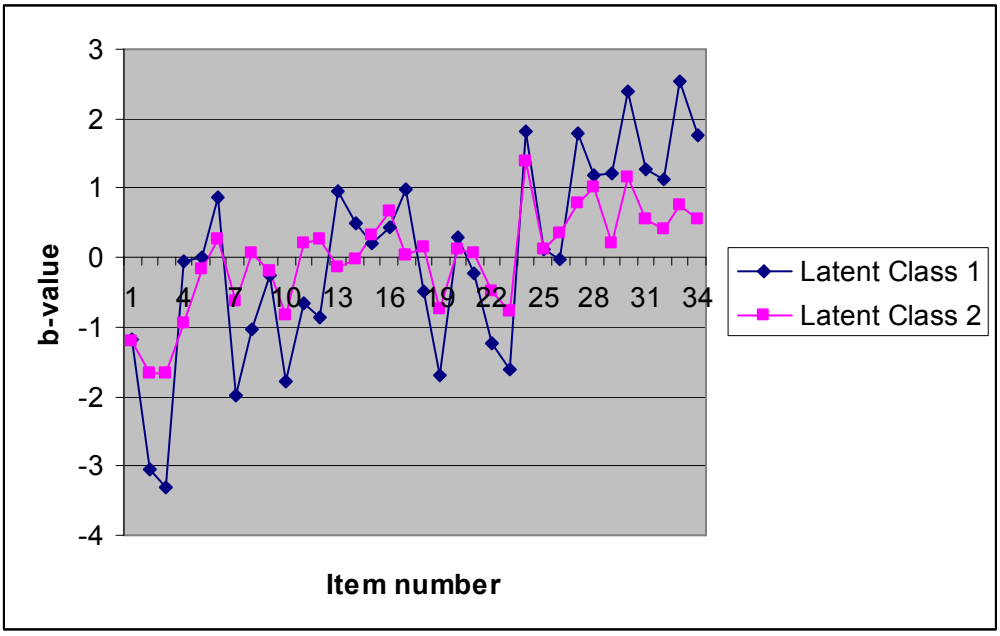


FIGURE 18: Item difficulties as a function of latent class

Based on the patterns of item difficulties for the two latent classes one could suggest several explanations. For example, it is possible that the two classes are related to the type of reading instruction employed in the students' classrooms. English language learners taught using a top-down, concept-driven approach (Weaver, 1994) would receive instruction through literacy activities, going from concepts to words. Students taught

more traditionally, in a bottom-up sequence (Weaver, 1994) would learn through phonics and word recognition. We would expect students taught using the latter approach to find items dealing with reasoning to be much more difficult and those requiring word recognition easier than students taught using the former approach.

Alternatively, membership in a latent class could have to do with the cognitive style of the student rather than the classroom. Some students may feel more comfortable learning English by memorizing words, while others might read books, newspapers and magazines. We would expect learners who memorize to find the questions on the first part of the test, that tend to require recall, to be extremely easy, and those items dealing with reading comprehension much more difficult. It would not be surprising if Asian students in particular utilized the former strategy since they may have learned to read thousands of characters in their native language. The fact that 83% of the Asian students were in the first class provides further evidence that this explanation may be tenable.

Again, this is not meant to be a comprehensive list of potential explanations for the patterns of item difficulties shown. Instead, these possible explanations serve two purposes. First, they model the types of explanations experts might posit in response to these patterns, and second they provide examples that can be carried further in the next step of this process.

The final step was to gain further evidence to support the hypotheses generated regarding the classes using covariates to predict latent class membership. The choice of these covariates may stem directly from step three of this process in a confirmatory manner, or may be more exploratory if step three yields few clues to why examinees perform differently on certain items.

Continuing with the previous example, if the cognitive style of the student was to be investigated, one covariate of interest might be whether or not the English language learner was born in the United States. One would expect students born outside of the U.S. to be more likely to learn to read English by memorizing words in an attempt to 'catch up' with their peers. Another possible covariate could deal with the type of instruction the student received in school. Students in dual language programs, which contain students who speak English learning Spanish (as an example) and Spanish speaking students learning English, might be more likely to learn their second language through literacy activities and not memorization.

Although information on the country in which the student was born and the type of ESL (English as a second language) program in which the student was enrolled were collected, only the former had enough data to be used as a covariate. Years of ESL instruction and grade were used in addition to birth country (US or not) as covariates. Unfortunately, none of these variables were significant predictors of latent class membership.

Latent Class versus Mantel-Haenszel Results

When comparing the results from the latent class analyses to those from the Mantel-Haenszel procedure we see most of the items identified using the manifest strategy were a subset of those identified using the latent approach. Two items, #9 and #25, were exceptions in that they showed manifest but not latent DIF. Item #9 (MH-χ^2 = 5.407, p = 0.020, ln(odds) = 0.452) had a relatively small amount of gender DIF, but item #25 (MH-χ^2 = 8.096, p = 0.004, ln(odds) = 0.995) had the most DIF of the items that function differently for the two ethnic groups. As Skaggs and Lissitz (1992, pg. 228)

noted, DIF "detection methods are not particularly reliable and [that] many of the identifications of biased items are statistical fluctuations of the item response data". Evidence that differential function was indeed an anomalous finding for item #25 comes from examining the deciles used for the Mantel-Haenszel χ^2. We see from these that one of the deciles accounts for more than half of the chi-square value. That decile, containing only 9 of the Asian students, seems aberrant in that the percentage of Asian students getting that item right was lower than in the previous decile with lower ability students. Had only 3 of the Asian students answering incorrectly gotten the item right the χ^2 value would have been cut in half. Given the seeming instability of this estimate it seems possible that this finding was indeed an anomaly. See Appendix C for the deciles for item #25.

The question that arises is why does only a subset of the items showing latent DIF get identified using a manifest approach? Inspection of the items functioning the most differently between the two latent classes provides some insights. The three items exhibiting the largest amount of positive DIF – #33 ($\Delta b = 1.787$), #30 ($\Delta b = 1.253$), and #34 ($\Delta b = 1.215$) – were all identified as functioning differentially using the Mantel-Haenszel approach. This was due to the differences in the percentages of Hispanic and Asian examinees, and males and females in the latent classes. For example, the first class consisted of 83% of the females and 66% of the males. We see from these numbers that while most of the males and females behaved alike on these test items, there is a relatively small percentage that responded differently. Since the examinees were so similar, the only way to see the very small differences between them was with items with an extremely large amount of DIF.

Theoretically, the same argument should hold for items with large amounts of negative latent DIF favoring the examinees in the first class. For these data, that is not the case. Of item #3 ($\Delta b = -1.665$), #2 ($\Delta b = -1.384$), and #7 ($\Delta b = -1.367$), only #7 showed manifest DIF. In this case that may be due to the item difficulties within the latent classes. Examination of Figure 18 on page 69 shows that in the first latent class the difficulties for items #2 and #3 were lower than -3.0, meaning these were extremely easy items. For items that easy, virtually all of the examinees in the first class will respond correctly, meaning that there will be no differences between males and females from that class on that item making it impossible to see the small differences between males and females that existed.

Summary

The question that remains is what was learned from a latent class analysis of the DIF that would be missed using a manifest approach? These analyses indicated that the relatively small number of items identified using the Mantel-Haenszel procedure, are an uninformative subset of the items functioning differentially from a latent perspective. That means they can provide only limited information regarding the true nature of the DIF. The latent class DIF analysis yielded many more insights into why the items functioned differentially.

For example, looking at the items identified by the two strategies, we see that the manifest approach with regard to gender pointed to two items, as did the same approach for ethnicity. The latent approach, however, identified all six of the items corresponding to the final passage on the test as having DIF. In this case, the manifest approach would

have test manufacturers examining individual items while the latent approach would yield an investigation at the level of the passage.

The latent DIF approach also provided the following insights into the nature of the educational advantage attribute underlying the DIF.

1. Mean abilities for the first latent class were higher than for the second class.
2. Items at the beginning of the test tended to be much easier and those at end of the test tended to be much harder for students in the first latent class.
3. Country of birth (United States or not), grade level, and years in ESL programs were not predictive of latent class membership.

Though these pieces of evidence did not provide an obvious answer as to what the latent classes were, they did yield some hints. Experts in literacy or linguistics might find the trends noted here fit neatly into the existing knowledge base, or they might be able to see patterns that were not clear to the untrained eye.

The bottom line is not that the procedure delineated in this research makes it a simple matter to find the cause of DIF. Rather, it conceptualizes the problem appropriately, and in doing so provides more information about which items function differentially and why.

Chapter 7: Discussion

The central premise of this research is that using manifest groups in DIF analyses is ill advised. Distinctions based on external characteristics of examinees are not helpful because the groups that result are neither homogeneous nor cognitively meaningful. Instead, DIF analyses should focus on what Dorans and Holland (1993) called an educational advantage attribute. By examining the latent dimensions underlying student performance it may be possible to identify and interpret the reasons behind differential item functioning.

Although this research focused on identifying DIF using a latent class perspective, it was important to retain the manifest distinctions most often used, and to interweave these with the latent classifications. This was advantageous for many reasons. First, mapping the manifest groups onto the latent classes provided a visual reference as to what the latent classes looked like – an important feature when trying to discuss latent DIF with those outside the psychometric community. Second, though it may be clear that an educational advantage attribute causes DIF, there will be political pressure to continue to think about DIF in terms of manifest groups. Whatever the reason for this pressure, it is nonetheless real and it cannot be ignored. Finally, examining the proportions of examinees from each of the manifest groups in the latent classes may help in conceptualizing the reasons items are functioning differentially.

Implications of this Research

To examine the issues arising due the heterogeneity of the manifest groups, this research simulated a variety of conditions under which the overlap between the latent

classes and manifest groups were manipulated. Results showed that power was affected by: (1) the amount of overlap between latent classes and manifest groups; (2) the magnitude of the DIF; (3) differences in the ability distributions of the latent classes; and (4) sample size considerations overall and in terms of the individual manifest groups. Additionally, it was found that as the overlap decreased (i.e. the latent classes became more mixed in terms of the manifest groups), the estimates of the magnitude of the DIF, as judged by the ln(odds), got increasingly worse, making it progressively more difficult to classify items as having problematic amounts of DIF. Finally, it was shown that the numbers of items incorrectly identified as functioning differentially was impacted by sample size, degree of contamination of the matching criterion, the amount of overlap, and the manifest proportions.

The problems surrounding the use of manifest groups that are combinations of examinees from different latent classes become even more pronounced when one considers the data from the test of English language proficiency used in this research. The simulated data used in the first stages of this study were very simplistic in that the majority of examinees in each manifest group came from different latent classes. The real data proved to be much more confusing in that the majority of examinees from each of the manifest groups were in the first latent class. That means that examinees in the manifest groups were more alike than different in terms of the dimensions underlying latent class membership. Though making generalizations to all assessments from a test of English language proficiency may be somewhat problematic, there does appear to be evidence that all of the issues raised in this research would actually worsen when applied to operational tests.

One implication of the simulation study performed is that the sample sizes typically used in DIF analyses appear to be too small. Examination of the sample sizes used by ETS (see Figure 19) shows that n's as low as 500 are acceptable during some parts of the test construction process. Based on this research it appears that when the sample size is that small, one will only see true DIF if the overlap between the latent class and manifest groups is greater than 80% and the magnitude of the DIF is large.

Sample sized used by ETS		
Smaller group n	Total n	When used
100	500	In test assembly
200	600	After administration but before scores reported
500		After score reporting

FIGURE 19: Sample sizes recommended by ETS

This research demonstrated a strategy for identifying the educational advantage attribute underlying student performance. Though a lack of substantive knowledge and appropriate covariates precluded the researcher from actually identifying the two latent classes, the strategy highlighted showed several advantages over typical manifest DIF approaches. First, examining the same data using the Mantel-Haenszel approach and the mixed Rasch model for detecting DIF, showed a wide discrepancy in the number of items identified as functioning differentially. If we assume the items identified using the manifest approach are a subset of the items with DIF due to some latent attribute, then it is clear that we have less information regarding the cause of the DIF with that manifest strategy. This will, in many cases, preclude us from ever determining the cause of the differential functioning.

At a more fundamental level, it appears that the DIF uncovered by traditional approaches may be attributable to differences in a relatively small number of examinees. The figure below depicts a scenario similar to the one uncovered in this research. In this case, the males in the first latent class, shown by the yellow portion of the top box, and a large portion of the females in that class should perform similarly on test items. Females in the second class, shown by the green portion of the bottom box, and a portion of the males in that class will also perform similarly. That leaves only the remaining males in the second class and the remaining females in the first class who will respond to the test items differently.

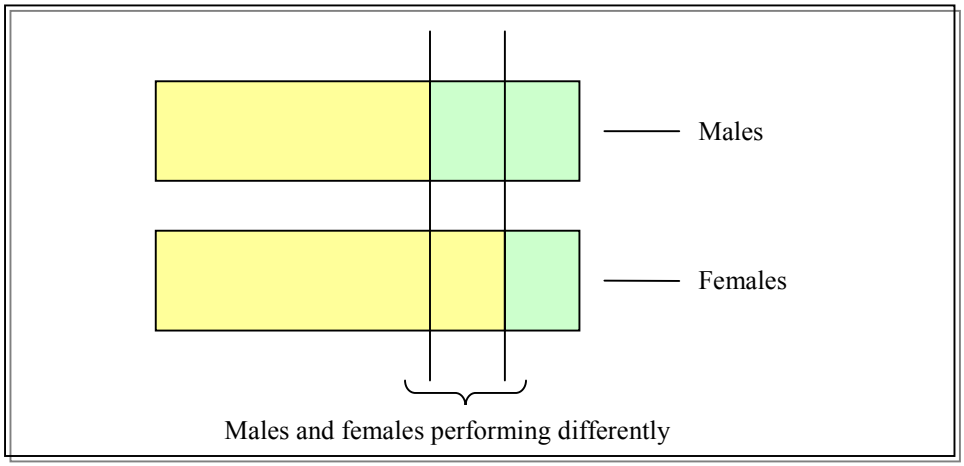


FIGURE 20: Mapping manifest distinctions onto a latent class analysis

The fact that the manifest DIF detected by traditional strategies may be an artifact of differences between relatively small numbers of examinees calls the appropriateness of those procedures into question. Do we really care if 20% of males and females respond differently to items if the remaining 80% respond in a similar fashion? The answer to that question is beyond the scope of this research. However, providing a mechanism that highlights the magnitude of the latent differences between manifest groups is one of the strengths of the strategy described in this paper.

Next Steps

This research used a test of English language proficiency to investigate an application of the mixed Rasch model in detecting DIF. The procedure delineated in this paper retained the typical manifest groups of gender and ethnicity and mapped those onto the latent classes. The findings indicated that a two class model held and that students in the two latent classes were a made up of examinees from all manifest groups. Though this was an atypical application for a DIF study in that Asians and Hispanics were the reference and focal groups, DIF across language groups has been previously examined (Chen & Henning, 1985; Kim, 2001). Future studies should apply this approach in more traditional applications, such as content area tests. In those applications it would be interesting to see if the same latent classes result when different pairs of manifest groups are included. That is, would the items identified as having latent DIF be the same in an analysis of whites and Hispanics as in a parallel analysis of whites and Asians? If so, more information would be gained with each successive pair of manifest groups tested.

Future studies should also be undertaken that utilize more complex models. This research was a first step in applying a latent approach to a situation in which manifest strategies are the norm. For that reason it seemed prudent to use the Rasch model. Additionally, since the Mantel-Haenszel procedure was utilized as a representative manifest strategy, the mixed Rasch model seemed a parallel procedure. Certainly a 2-parameter logistic model could be employed for multiple-choice items, however interpretations of differences between latent groups with regard to the item parameters become more difficult. A 3-PL model might prove even more problematic given the inherent difficulty in estimating the pseudo-guessing parameter.

Latent class DIF studies that build upon this research need to incorporate differential test speededness into the model. For this study, the final ten items were dropped to 'ensure' that test speededness would not impact the findings. In reality, students in different latent classes may have different patterns of missingness, some due to not reaching items, others due to judiciously skipping high difficulty items to spend the limited time on other items. To account for this a model like Yamamoto and Everson's adaptation of the hybrid model (1996) should be incorporated into the mixed Rasch model. That way some of a respondent's answers could be based on the ability and latent class, and for the remaining items, where random guessing occurred, a multinomial model would hold.

Given that the sources of item difficulty differ widely from item to item (Whitely & Schneider, 1981), another important next step would be to model the elemental components within items rather than looking at the items as a whole. This sort of decomposition has been done for verbal items (Sheehan & Mislevy, 1990; Janssen & DeBoeck, 1997), mathematical problems (Fischer, 1973; Embretson, 1995), and nonverbal items (Green & Smith, 1987, Embretson, 1998). One way to model item difficulty using elemental components is to replace the item difficulty term in the Rasch model with the sum of the products of the scored features of the items and the weights of those features, plus a constant. Using Embretson's (1998) model of abstract reasoning as an example, one could have five scored variables (number of rules, abstract correspondence, distortion, fusion, and overlay), and each of those variables would have a weight associated with it. For an item with three rules and none of the other features, the item difficulty would be three times the weight for the number of rules variable plus a

constant. Taking this example further, it seems possible that items would function differentially because individuals in a latent class had trouble with items with abstraction. It should be noted that using some sort of elemental components approach would provide researchers with more information regarding the underlying cause of DIF; however this is not a post hoc strategy like many others. Items used on an assessment would need to be 'model-able' using a set of elemental components, and would therefore need to be chosen for the test with that in mind.

Conclusion

Strategies for the detection of DIF have evolved over the past several decades; however most have had one inherent flaw – using manifest groups as the basis from which to make comparisons. The pithy comment of Skaggs and Lissitz (1992, pg. 239), that "*Black* is not a cognitively meaningful dimension and not even a well-defined one for that matter" sums up the difficulties with those sorts of approaches. This paper has detailed a latent class strategy for DIF detection that retains the manifest groups often considered, but does so within a latent framework. While some might argue the specifics of the approach, it is hoped that the general premise is incontestable.

Taking a latent approach is clearly not the easy road to travel. Test manufacturers will argue because DIF analyses would be more time consuming using latent strategies. State and local agencies will be resistant because they will have problems explaining these sorts of analyses to stakeholders who may have troubling grasping the idea of latent constructs. Those concerns are real and have merit; however it is time to convince all concerned about equity in testing that what is gained is worth it. Most importantly, using a latent class perspective, we base individual differences in human behavior on

potentially meaningful dimensions rather than external characteristics. Once that happens, we gain the possibility of actually explaining why items function differentially. These benefits clearly seem worth the effort.

Appendix A: Results of Simulation Study

50/50 manifest split, 2000 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.435	0.420	0.420	0.417	0.401	0.406
90%	0.350	0.330	0.335	0.310	0.320	0.305
80%	0.260	0.250	0.252	0.227	0.241	0.217
70%	0.170	0.150	0.163	0.158	0.154	0.143
60%	0.085	0.085	0.085	0.073	0.073	0.073

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.845	0.870	0.840	0.840	0.808	0.821
90%	0.680	0.645	0.672	0.642	0.645	0.618
80%	0.505	0.465	0.503	0.467	0.483	0.444
70%	0.340	0.305	0.330	0.303	0.318	0.285
60%	0.165	0.160	0.168	0.140	0.159	0.138

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	1.280	1.295	1.272	1.282	1.242	1.250
90%	0.995	0.950	1.005	0.958	0.978	0.937
80%	0.735	0.685	0.740	0.685	0.718	0.661
70%	0.485	0.445	0.492	0.445	0.478	0.435
60%	0.250	0.220	0.243	0.220	0.234	0.210

80/20 manifest split, 2000 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.435	0.430	0.422	0.423	0.395	0.401
90%	0.335	0.330	0.335	0.315	0.311	0.303
80%	0.260	0.240	0.253	0.240	0.236	0.208
70%	0.170	0.175	0.163	0.142	0.154	0.144
60%	0.090	0.085	0.080	0.075	0.070	0.059

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.855	0.840	0.850	0.835	0.825	0.824
90%	0.680	0.635	0.658	0.640	0.639	0.604
80%	0.505	0.460	0.502	0.452	0.486	0.439
70%	0.325	0.285	0.340	0.300	0.318	0.288
60%	0.175	0.145	0.162	0.150	0.151	0.130

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	1.270	1.270	1.278	1.272	1.241	1.243
90%	0.980	0.940	1.003	0.953	0.977	0.924
80%	0.725	0.680	0.737	0.677	0.718	0.659
70%	0.490	0.450	0.490	0.457	0.471	0.424
60%	0.260	0.240	0.233	0.227	0.242	0.214

50/50 manifest split, 500 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.425	0.450	0.422	0.420	0.415	0.399
90%	0.375	0.335	0.317	0.330	0.335	0.325
80%	0.275	0.250	0.258	0.235	0.248	0.224
70%	0.160	0.150	0.170	0.140	0.151	0.142
60%	0.085	0.105	0.077	0.068	0.081	0.066

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.860	0.855	0.848	0.838	0.828	0.831
90%	0.665	0.650	0.680	0.643	0.665	0.614
80%	0.515	0.435	0.500	0.467	0.488	0.450
70%	0.335	0.295	0.333	0.305	0.325	0.288
60%	0.145	0.150	0.175	0.153	0.154	0.130

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	1.285	1.290	1.267	1.283	1.241	1.254
90%	1.015	0.975	0.993	0.967	0.992	0.940
80%	0.735	0.715	0.742	0.678	0.727	0.663
70%	0.475	0.460	0.482	0.457	0.474	0.432
60%	0.240	0.190	0.255	0.210	0.235	0.214

80/20 manifest split, 500 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.425	0.460	0.408	0.422	0.391	0.395
90%	0.340	0.335	0.323	0.313	0.313	0.303
80%	0.275	0.220	0.253	0.228	0.236	0.234
70%	0.175	0.120	0.175	0.157	0.161	0.141
60%	0.060	0.085	0.092	0.058	0.078	0.072

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	0.870	0.875	0.852	0.873	0.831	0.807
90%	0.635	0.630	0.648	0.632	0.644	0.615
80%	0.485	0.480	0.517	0.438	0.484	0.441
70%	0.325	0.315	0.337	0.280	0.300	0.275
60%	0.160	0.130	0.157	0.138	0.176	0.124

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	1.295	1.310	1.270	1.288	1.253	1.245
90%	1.010	0.920	1.013	0.952	0.987	0.959
80%	0.725	0.670	0.758	0.682	0.717	0.660
70%	0.485	0.440	0.487	0.433	0.485	0.438
60%	0.220	0.205	0.262	0.243	0.240	0.202

50/50 manifest split, 2000 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	97.5	95.5	97.7	94.7	92.5	88.0
90%	88.5	83.0	86.7	75.2	77.5	68.0
80%	72.5	59.0	66.7	49.7	54.7	43.9
70%	36.0	25.0	31.0	27.8	27.1	22.8
60%	10.5	13.5	12.8	8.8	9.3	8.5

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	100.0	100.0	100.0	100.0	100.0	100.0
90%	100.0	100.0	100.0	99.8	100.0	99.7
80%	100.0	100.0	99.7	97.7	98.5	95.8
70%	93.5	78.0	90.7	78.7	81.3	69.2
60%	34.5	30.5	31.8	24.3	27.3	21.7

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	100.0	100.0	100.0	100.0	100.0	100.0
90%	100.0	100.0	100.0	100.0	100.0	100.0
80%	100.0	100.0	100.0	100.0	100.0	100.0
70%	100.0	98.5	99.7	98.7	99.7	97.2
60%	68.5	52.0	66.5	49.3	58.5	45.0

80/20 manifest split, 2000 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	91.5	83.5	86.8	83.2	80.0	71.1
90%	75.5	64.5	69.0	59.3	59.1	51.6
80%	50.0	40.0	47.8	37.8	37.2	28.3
70%	25.5	22.5	22.2	16.0	18.0	14.4
60%	10.0	8.0	8.3	7.2	8.3	8.4

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	100.0	100.0	100.0	100.0	100.0	99.9
90%	100.0	100.0	99.8	99.8	99.6	96.8
80%	98.0	89.5	96.3	89.7	94.3	84.3
70%	73.5	52.5	74.3	57.5	62.5	51.3
60%	25.0	19.5	24.7	16.0	18.6	13.3

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	100.0	100.0	100.0	100.0	99.9	100.0
90%	100.0	100.0	100.0	100.0	100.0	100.0
80%	100.0	99.5	100.0	100.0	99.9	99.4
70%	95.5	91.0	96.3	92.7	95.8	84.6
60%	51.5	42.5	42.5	36.7	41.2	31.3

50/50 manifest split, 500 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	50.0	44.0	46.5	38.0	40.1	31.3
90%	36.0	30.5	28.3	24.3	27.0	21.6
80%	19.5	13.5	19.2	14.3	15.5	12.1
70%	12.0	7.5	10.0	6.5	8.1	8.5
60%	5.5	7.5	5.7	4.7	3.8	5.4

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	98.0	93.5	96.7	93.0	93.2	88.6
90%	88.0	79.0	88.5	77.7	81.2	68.0
80%	65.5	48.5	61.3	49.3	54.8	44.3
70%	35.0	22.0	29.7	25.3	27.0	19.8
60%	8.0	4.5	11.7	9.5	8.1	7.2

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	100.0	100.0	100.0	100.0	98.4	99.7
90%	99.5	99.0	99.3	98.5	98.1	96.6
80%	95.0	90.5	91.7	86.0	87.5	77.4
70%	61.0	52.5	60.5	49.7	55.1	43.3
60%	15.5	12.0	20.7	12.8	14.7	12.8

80/20 manifest split, 500 examinees

	2 items, DIF=0.40		6 items, DIF=0.40		10 items, DIF=0.40	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	34.0	32.0	29.7	26.5	22.3	19.4
90%	20.0	17.0	20.2	16.8	16.6	14.8
80%	16.5	10.0	14.0	10.3	12.8	9.7
70%	9.5	4.5	8.5	7.0	7.3	5.6
60%	5.0	5.0	4.8	5.3	4.7	4.4

	2 items, DIF=0.80		6 items, DIF=0.80		10 items, DIF=0.80	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	86.0	85.0	86.8	81.3	79.2	69.3
90%	61.0	56.0	63.0	54.3	59.2	49.0
80%	39.0	36.0	47.8	29.0	37.6	28.7
70%	24.0	16.5	22.2	15.0	17.1	13.7
60%	10.5	3.0	6.5	6.2	7.8	6.0

	2 items, DIF=1.20		6 items, DIF=1.20		10 items, DIF=1.20	
	No Shift	Shift	No Shift	Shift	No Shift	Shift
100%	99.0	98.5	100.0	97.7	97.3	95.7
90%	95.0	87.5	97.3	88.7	92.0	88.0
80%	77.0	63.5	77.0	64.7	72.5	59.7
70%	41.0	33.0	41.5	33.5	39.0	30.5
60%	7.0	8.0	12.5	11.8	13.1	9.2

Appendix B: BUGS output for simulated data

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.7352	0.1012	0.003950	0.5360	0.7400	0.9200
Propfemale[2]	0.2648	0.1012	0.003950	0.0800	0.2600	0.4640
Propmale[1]	0.2500	0.1062	0.004167	0.0600	0.2440	0.4640
Propmale[2]	0.7500	0.1062	0.004167	0.5360	0.7560	0.9400
bdif[1]	0.0465	0.5188	0.008993	-1.0300	0.0651	1.0140
bdif[2]	-0.6747	0.4462	0.007314	-1.6030	-0.6592	0.1560
bdif[3]	0.6352	0.342	0.004997	-0.0477	0.6376	1.2930
bdif[4]	0.5610	0.3285	0.004719	-0.0878	0.5629	1.1970
bdif[5]	**0.7124**	**0.3078**	**0.004237**	**0.1066**	**0.7135**	**1.3150**
bdif[6]	0.0817	0.3184	0.004465	-0.5688	0.0885	0.6934
bdif[7]	-0.5418	0.3258	0.004616	-1.2050	-0.5363	0.0835
bdif[8]	**-1.1240**	**0.3678**	**0.006065**	**-1.8760**	**-1.1160**	**-0.4242**
bdif[9]	-0.5972	0.3383	0.005133	-1.2770	-0.5924	0.0521
bdif[10]	0.5209	0.3246	0.004957	-0.1254	0.5247	1.1450
bdif[11]	0.1842	0.3513	0.004722	-0.5276	0.1906	0.8572
bdif[12]	0.1334	0.3368	0.004856	-0.5440	0.1397	0.7798
bdif[13]	-0.1715	0.3310	0.004854	-0.8498	-0.1637	0.4529
bdif[14]	**-0.7650**	**0.3653**	**0.005984**	**-1.5070**	**-0.7577**	**-0.0680**
bdif[15]	0.0922	0.3211	0.004420	-0.5457	0.0952	0.7158
bdif[16]	0.1612	0.3163	0.004484	-0.4691	0.1626	0.7719
bdif[17]	0.4244	0.3280	0.004467	-0.2314	0.4255	1.0560
bdif[18]	-0.1761	0.3506	0.005006	-0.8820	-0.1717	0.4983
bdif[19]	0.3762	0.3700	0.005588	-0.3655	0.3800	1.0870
bdif[20]	0.1213	0.4670	0.075290	-0.8457	0.1490	0.9372
mu[1]	-0.0757	0.1349	0.004399	-0.3142	-0.0853	0.2126
mu[2]	-0.9446	0.1331	0.004213	-1.2270	-0.9374	-0.7061

Table B1. Statistics for 70% overlap on 6 items with different ability distributions (500 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.2102	0.1228	0.005186	0.0120	0.2000	0.4600
Propfemale[2]	0.7898	0.1228	0.005186	0.5400	0.8000	0.9880
Propmale[1]	0.7337	0.1189	0.004884	0.5160	0.7360	0.9520
Propmale[2]	0.2663	0.1189	0.004884	0.0480	0.2640	0.4840
bdif[1]	-0.6298	0.4206	0.006023	-1.4520	-0.6320	0.2080
bdif[2]	0.4502	0.4181	0.005962	-0.3592	0.4441	1.2910
bdif[3]	**-0.7814**	**0.3368**	**0.005033**	**-1.4430**	**-0.7830**	**-0.1204**
bdif[4]	-0.2506	0.3121	0.004285	-0.8615	-0.2513	0.3675
bdif[5]	-0.3710	0.3383	0.005348	-1.0290	-0.3713	0.2996
bdif[6]	-0.2039	0.3206	0.004642	-0.8049	-0.2141	0.4582
bdif[7]	**1.1030**	**0.3668**	**0.005820**	**0.4235**	**1.0890**	**1.8640**
bdif[8]	0.1390	0.3546	0.007231	-0.5513	0.1383	0.8431
bdif[9]	**0.8093**	**0.3805**	**0.006530**	**0.1038**	**0.7924**	**1.6070**
bdif[10]	0.1498	0.3750	0.006811	-0.5538	0.1374	0.9265
bdif[11]	-0.3058	0.3731	0.005256	-1.0380	-0.3074	0.4305
bdif[12]	-0.1111	0.3681	0.006249	-0.8279	-0.1109	0.6217
bdif[13]	-0.0387	0.3225	0.004529	-0.6480	-0.0463	0.6199
bdif[14]	-0.1358	0.3075	0.004090	-0.7235	-0.1415	0.4851
bdif[15]	0.3872	0.3626	0.007575	-0.2948	0.3749	1.1370
bdif[16]	0.0625	0.3168	0.004731	-0.5426	0.0558	0.7086
bdif[17]	0.0908	0.3220	0.004599	-0.5258	0.0848	0.7380
bdif[18]	0.1780	0.3566	0.005731	-0.4874	0.1666	0.9140
bdif[19]	0.0407	0.4140	0.007708	-0.7357	0.0265	0.8964
bdif[20]	-0.5825	0.4253	0.005220	-1.4100	-0.5820	0.2597
mu[1]	-0.8362	0.1143	0.003473	-1.0770	-0.8294	-0.6308
mu[2]	-0.1169	0.1429	0.004851	-0.3689	-0.1260	0.1800

Table B2. Statistics for 80% overlap on 6 items with different ability distributions (500 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.3860	0.0617	0.002609	2.2700	0.3850	0.4950
Propfemale[2]	0.6140	0.0617	0.002609	0.5050	0.6150	0.7300
Propmale[1]	0.5944	0.0595	0.002516	0.4970	0.5900	0.7100
Propmale[2]	0.4056	0.0595	0.002516	0.2900	0.4100	0.5030
bdif[1]	-0.1068	0.2666	0.004916	-0.6188	-0.1112	0.4249
bdif[2]	0.0977	0.2596	0.004861	-0.3885	0.0899	0.6327
bdif[3]	**-0.7528**	**0.1860**	0.002716	**-1.1140**	**-0.7533**	**-0.3848**
bdif[4]	**-0.9529**	**0.1694**	0.002483	**-1.2830**	**-0.9526**	**-0.6198**
bdif[5]	**-0.8477**	**0.1663**	0.002494	**-1.1710**	**-0.8493**	**-0.5178**
bdif[6]	0.0658	0.1840	0.002928	-0.2835	0.0621	0.4401
bdif[7]	**0.5119**	**0.2170**	0.004197	**0.1004**	**0.5060**	**0.9485**
bdif[8]	**0.8532**	**0.2484**	0.005617	**0.3852**	**0.8460**	**1.3560**
bdif[9]	**0.8306**	**0.2288**	0.004567	**0.3968**	**0.8262**	**1.2910**
bdif[10]	-0.4427	0.1971	0.003798	-0.8230	-0.4459	-0.0502
bdif[11]	0.0145	0.2134	0.003839	-0.3927	0.0082	0.4452
bdif[12]	-0.2942	0.1942	0.003289	-0.6640	-0.2979	0.0931
bdif[13]	0.0308	0.2002	0.003577	-0.3465	0.0261	0.4384
bdif[14]	0.3462	0.1991	0.003675	-0.0345	0.3431	0.7464
bdif[15]	-0.1126	0.1829	0.003005	-0.4666	-0.1147	0.2517
bdif[16]	-0.1088	0.1815	0.002994	-0.4600	-0.1113	0.2564
bdif[17]	0.3772	0.2174	0.004149	-0.0360	0.3736	0.8121
bdif[18]	-0.1696	0.1938	0.003029	-0.5429	-0.1726	0.2203
bdif[19]	0.2849	0.2443	0.005112	-0.1678	0.2760	0.7932
bdif[20]	0.3752	0.3164	0.006529	-0.1843	0.3528	1.0650
mu[1]	-1.0150	0.0831	0.003128	-1.1860	-1.0100	-0.8657
mu[2]	-0.0109	0.0799	0.002870	-0.1574	-0.0127	0.1501

Table B3. Statistics for 70% overlap on 6 items with different ability distributions (2000 examinees)

node	mean	sd	MC error	2.5%	median	97.5%
Propfemale[1]	0.2012	0.4972	0.001971	0.1090	0.1990	0.3050
Propfemale[2]	0.7988	0.4972	0.001971	0.6950	0.8010	0.8910
Propmale[1]	0.8154	0.4511	0.001748	0.7260	0.8160	0.9010
Propmale[2]	0.1846	0.4511	0.001748	0.0990	0.1840	0.2740
bdif[1]	0.2562	0.2209	0.003359	-0.1595	0.2485	0.7113
bdif[2]	0.1158	0.1907	0.002316	-0.2486	0.1120	0.4977
bdif[3]	**-0.9857**	**0.1535**	**0.001907**	**-1.2870**	**-0.9854**	**-0.6833**
bdif[4]	**-0.9149**	**0.1417**	**0.001781**	**-1.1940**	**-0.9146**	**-0.6379**
bdif[5]	**-0.7994**	**0.1334**	**0.001460**	**-1.0630**	**-0.7993**	**-0.5371**
bdif[6]	0.2648	0.1454	0.001796	-0.0170	0.2640	0.5508
bdif[7]	**0.5962**	**0.1499**	**0.001709**	**0.3056**	**0.5954**	**0.8942**
bdif[8]	**0.9360**	**0.1609**	**0.002121**	**0.6285**	**0.9335**	**1.2600**
bdif[9]	**0.9431**	**0.1603**	**0.002035**	**0.6357**	**0.9408**	**1.2640**
bdif[10]	-0.0665	0.1501	0.001694	-0.3591	-0.0668	0.2300
bdif[11]	0.2824	0.1697	0.002072	-0.0462	0.2803	0.6216
bdif[12]	0.1423	0.1577	0.001836	-0.1639	0.1420	0.4534
bdif[13]	-0.0725	0.1435	0.001761	-0.3526	-0.0727	0.2108
bdif[14]	0.1351	0.1405	0.001552	-0.1366	0.1343	0.4142
bdif[15]	-0.0718	0.1369	0.001424	-0.3391	-0.0723	0.1983
bdif[16]	-0.1932	0.1419	0.001738	-0.4727	-0.1926	0.0846
bdif[17]	0.1456	0.1500	0.001780	-0.1494	0.1453	0.4402
bdif[18]	-0.2427	0.1622	0.002447	-0.5553	-0.2453	0.0804
bdif[19]	-0.2657	0.1688	0.002221	-0.5942	-0.2661	0.0674
bdif[20]	-0.2049	0.1914	0.002020	-0.5736	-0.2067	0.1744
mu[1]	-1.0060	0.0654	0.002364	-1.1510	-1.0030	-0.8777
mu[2]	0.0428	0.0629	0.002014	-0.0768	0.0416	0.1687

Table B4. Statistics for 80% overlap on 6 items with different ability distributions (2000 examinees)

Appendix C: Mantel-Haenszel Deciles for Item #25

Decile	Item Score	Group		Total
		Focal	Reference	
1	0	5	67	72
	1	0	27	27
	Total	5	94	99
2	0	9	55	64
	1	1	27	28
	Total	10	82	92
3	0	5	55	60
	1	3	59	62
	Total	8	114	122
4	0	6	17	23
	1	3	50	53
	Total	9	67	76
5	0	5	18	23
	1	12	105	117
	Total	17	123	140
6	0	1	2	3
	1	9	44	53
	Total	10	46	56
7	0	0	6	6
	1	20	117	137
	Total	20	123	143
8	0	0	1	1
	1	12	52	64
	Total	12	53	65
9	0	0	0	0
	1	32	129	161
	Total	32	129	161
10	0	0	0	0
	1	13	49	62
	Total	13	49	62

Appendix D: GAUSS Code

```
NEW;
RNDSEED(3905482);
LC11=400;
LC21=400-LC11;
OVERLAP=1.0;
SHIFT=0.0;
NR=100;
NI=20;
CHI=ZEROS(NI,NR+1);
SIGNIF=ZEROS(NI,NR+1);
ODDS=ZEROS(NI,NR+1);
R=1;
DO WHILE R<=NR;

/* GENERATE DATA */

TNS=500;
NFC=400;
NSC=TNS-NFC;
RESP=ZEROS(TNS,NI+2);
X=ZEROS(TNS,NI+1);
RESP[.,1]=(RNDN(TNS,1));
Z=NFC+1;
DO WHILE Z<=TNS;
RESP[Z,1]=RESP[Z,1]-SHIFT;
Z=Z+1;
ENDO;
B=ZEROS(1,NI);
BB=ZEROS(1,NI);
B={2.0 1.6 1.2 0.8 0.4 0.0 -0.4 -0.8 -1.2 -1.6 -2.0 -1.6 -1.2 -0.8 -0.4 0.0 0.4 0.8 1.2 1.6};
BB={2.0 1.6 1.2 0.0 0.4 0.0 -0.4 0.0 -1.2 -1.6 -2.0 -1.6 -1.2 -0.8 -0.4 0.0 0.4 0.8 1.2 1.6};
J=1;
DO WHILE J<=NFC;
K=1;
DO WHILE K<=NI;
PROB = (EXP((RESP[J,1]-B[1,K])))/(1 + EXP((RESP[J,1]-B[1,K])));
A=RNDU(1,1);

IF A<=PROB;
RESP[J,K+1]=1;
ENDIF;
IF A>=PROB;
RESP[J,K+1]=0;
ENDIF;
K=K+1;
ENDO;
IF J <=LC11;
RESP[J,NI+2]=1;
ELSE;
```

```
RESP[J,NI+2]=0;
ENDIF;
J=J+1;
ENDO;
J=1;
DO WHILE J<=NSC;
K=1;
DO WHILE K<=NI;
PROB = (EXP((RESP[NFC+J,1]-BB[1,K])))/(1 + EXP((RESP[NFC+J,1]-BB[1,K])));
A=RNDU(1,1);
IF A<=PROB;
RESP[NFC+J,K+1]=1;
ENDIF;
IF A>=PROB;
RESP[NFC+J,K+1]=0;
ENDIF;
K=K+1;
ENDO;
IF J<=LC21;
RESP[NFC+J,NI+2]=1;
ELSE;
RESP[NFC+J,NI+2]=0;
ENDIF;
J=J+1;
ENDO;
X[.,.]=RESP[.,2:NI+2];

/* SEPARATE INTO MALE AND FEMALE MATRICES*/

MALE=ZEROS(400,NI+1);
FEMALE=ZEROS(100,NI+1);
I=0;
J=0;
N=1;
DO WHILE N<=TNS;
IF X[N,NI+1]==1;
I=I+1;
MALE[I,1:NI]=X[N,1:NI];
ELSEIF X[N,NI+1]==0;
J=J+1;
FEMALE[J,1:NI]=X[N,1:NI];
ENDIF;
N=N+1;
ENDO;
NUMFEM=J;
NUMMALE=I;
PROPFEM=J/(I+J);
PROPMALE=I/(I+J);

/* GET TOTAL SCORES FOR ALL EXAMINEES*/
```

```
N=1;
DO WHILE N<=400;
MALE[N,NI+1]=SUMC((MALE[N,1:NI])');
N=N+1;
ENDO;
N=1;
DO WHILE N<=100;
FEMALE[N,NI+1]=SUMC((FEMALE[N,1:NI])');
N=N+1;
ENDO;

/* CREATE FREQUENCY TABLES FOR EACH ITEM*/

FREQ=ZEROS(NI*(NI+1),13);
X=1;
DO WHILE X<=NI;
B=0;
DO WHILE B<=NI;
A=1;
FREQMR=0;
FREQMW=0;

FREQFR=0;
FREQFW=0;
DO WHILE A<=400;
IF MALE[A,NI+1]==B;
IF MALE[A,X]==1;
FREQMR=FREQMR+1;
ENDIF;
IF MALE[A,X]==0;
FREQMW=FREQMW+1;
ENDIF;
ENDIF;
A=A+1;
ENDO;
AA=1;
DO WHILE AA<=100;
IF FEMALE[AA,NI+1]==B;
IF FEMALE[AA,X]==1;
FREQFR=FREQFR+1;
ENDIF;
IF FEMALE[AA,X]==0;
FREQFW=FREQFW+1;
ENDIF;
ENDIF;
AA=AA+1;
ENDO;
FREQ[(X-1)*(NI+1)+B+1,1]=FREQMR;
FREQ[(X-1)*(NI+1)+B+1,2]=FREQFR;
FREQ[(X-1)*(NI+1)+B+1,3]=FREQMR+FREQFR;
FREQ[(X-1)*(NI+1)+B+1,4]=FREQMW;
```

```
FREQ[(X-1)*(NI+1)+B+1,5]=FREQFW;
FREQ[(X-1)*(NI+1)+B+1,6]=FREQMW+FREQFW;
FREQ[(X-1)*(NI+1)+B+1,7]=FREQMR+FREQMW;
FREQ[(X-1)*(NI+1)+B+1,8]=FREQFR+FREQFW;
FREQ[(X-1)*(NI+1)+B+1,9]=FREQMR+FREQFR+FREQMW+FREQFW;
IF FREQ[(X-1)*(NI+1)+B+1,9]==0;
FREQ[(X-1)*(NI+1)+B+1,10]=0;
ELSE;
FREQ[(X-1)*(NI+1)+B+1,10]=FREQ[(X-1)*(NI+1)+B+1,7]*FREQ[(X-
1)*(NI+1)+B+1,3]/FREQ[(X-1)*(NI+1)+B+1,9];
ENDIF;
IF FREQ[(X-1)*(NI+1)+B+1,9]==0;
FREQ[(X-1)*(NI+1)+B+1,11]=0;
ELSEIF FREQ[(X-1)*(NI+1)+B+1,9]==1;
FREQ[(X-1)*(NI+1)+B+1,11]=0;
ELSE;
FREQ[(X-1)*(NI+1)+B+1,11]=FREQ[(X-1)*(NI+1)+B+1,7]*FREQ[(X-
1)*(NI+1)+B+1,3]*FREQ[(X-1)*(NI+1)+B+1,8]*FREQ[(X-1)*(NI+1)+B+1,6]/((FREQ[(X-
1)*(NI+1)+B+1,9])*(FREQ[(X-1)*(NI+1)+B+1,9])*(FREQ[(X-1)*(NI+1)+B+1,9]-1));
ENDIF;
IF FREQ[(X-1)*(NI+1)+B+1,9]==0;
FREQ[(X-1)*(NI+1)+B+1,12]=0;
ELSE;
FREQ[(X-1)*(NI+1)+B+1,12]=FREQ[(X-1)*(NI+1)+B+1,1]*FREQ[(X-
1)*(NI+1)+B+1,5]/FREQ[(X-1)*(NI+1)+B+1,9];
ENDIF;
IF FREQ[(X-1)*(NI+1)+B+1,9]==0;
FREQ[(X-1)*(NI+1)+B+1,13]=0;
ELSE;
FREQ[(X-1)*(NI+1)+B+1,13]=FREQ[(X-1)*(NI+1)+B+1,2]*FREQ[(X-
1)*(NI+1)+B+1,4]/FREQ[(X-1)*(NI+1)+B+1,9];
ENDIF;
B=B+1;
ENDO;
X=X+1;
ENDO;

/* CHI SQUARE FOR EACH ITEM */

C=1;
DO WHILE C<=NI;
IF FREQ[(C-1)*(NI+1)+C,9]==0;
CHI[C,R]=0;
ELSE;
CHI[C,R]=(ABS(SUMC(FREQ[((C-1)*(NI+1)+1):((C-1)*(NI+1)+NI+1),1])-SUMC(FREQ[((C-
1)*(NI+1)+1):((C-1)*(NI+1)+NI+1),10]))-0.5)*(ABS(SUMC(FREQ[((C-1)*(NI+1)+1):((C-
1)*(NI+1)+NI+1),1])-SUMC(FREQ[((C-1)*(NI+1)+1):((C-1)*(NI+1)+NI+1),10]))-
0.5)/(SUMC(FREQ[((C-1)*(NI+1)+1):((C-1)*(NI+1)+NI+1),11]));
ENDIF;
IF CHI[C,R]>=3.84146;
SIGNIF[C,R]=1;
```

```
ENDIF;
CHI[C,NR+1]=(SUMC((CHI[C,1:NR])'))/NR;
SIGNIF[C,NR+1]=SUMC((SIGNIF[C,1:NR])');
ODDS[C,R]=LN(SUMC(FREQ[(C-1)*(NI+1)+1:(C-1)*(NI+1)+NI+1,12])/SUMC(FREQ[(C-
1)*(NI+1)+1:(C-1)*(NI+1)+NI+1,13]));
ODDS[C,NR+1]=(SUMC((ODDS[C,1:NR])'))/NR;
C=C+1;
ENDO;
R=R+1;
ENDO;
FORMAT/M1/RDN 8,2;
CHI[1:NI,NR+1];
SIGNIF[1:NI,NR+1];
ODDS[1:NI,NR+1];
END;
```

Appendix E: Annotated WINBUGS Code

```
model mixed_Rasch_model
{
probaf1 ~ dunif(0,1);     # Prior for the proportion of Asian females in the first latent class
probhf1 ~ dunif(0,1);     # Prior for the proportion of Hispanic females in the first latent class
probam1 ~ dunif(0,1);     # Prior for the proportion of Asian males in the first latent class
probhm1 ~ dunif(0,1);     # Prior for the proportion of Hispanic males in the first latent class

for (k in 1:2){
        muaf[k] ~ dnorm(0,1);           # Mean ability for Asian females
        muhf[k] ~ dnorm(0,1);           # Mean ability for Hispanic females
        muam[k] ~ dnorm(0,1);           # Mean ability for Asian males
        muhm[k] ~ dnorm(0,1);           # Mean ability for Hispanic females
        for (i in 1:N){
  thetaaf[i,k] ~ dnorm(muaf[k],1);
  thetahf[i,k] ~ dnorm(muhf[k],1);
  thetaam[i,k] ~ dnorm(muam[k],1);
  thetahm[i,k] ~ dnorm(muhm[k],1);
}
}
# x[i,46] is a dichotomous variable indicating gender (0=male, 1=female)
# x[i,47] is a dichotomous variable indicating ethnicity (0=Hispanic, 1=Asian)
# The probability of an examinee being in the first latent class is contingent upon the manifest
# characteristics of that examinee.
for (i in 1:N){
pi1[i] <- probaf1*x[i,46]*x[i,47]+ probhf1*(1-x[i,46])*x[i,47]+ probam1*x[i,46]*(1-
x[i,47])+probhm1*(1-x[i,46])*(1-x[i,47]);
  r[i] ~ dbern(pi1[i]);      # Trigger for latent class membership
}
for (j in 1:J-1){
        bdif[j] <- b[j,1]-b[j,2];           # DIF calculation
        for (k in 1:2){
                b[j,k] ~ dnorm(0,1);        # Prior for the item parameter in the latent classes
        }
}
b[J,1] <- -1*sum(b[1:(J-1),1]);             # Item difficulties within a class sum to zero
b[J,2] <- -1*sum(b[1:(J-1),2]);
bdif[J] <- b[J,1]-b[J,2];

# Rasch model with triggers for latent class membership and manifest characteristics
for (i in 1:N){
        for (j in 1:J){
        numer[i,j] <- ((exp((r[i])*((thetaaf[i,1]*x[i,46]*x[i,47]+ thetahf[i,1]*(1-x[i,46])*x[i,47]+
thetaam[i,1]*x[i,46]*(1-x[i,47])+thetahm[i,1]*(1-x[i,46])*(1-x[i,47]))-b[j,1])+(1-
r[i])*((thetaaf[i,2]*x[i,46]*x[i,47]+ thetahf[i,2]*(1-x[i,46])*x[i,47]+ thetaam[i,2]*x[i,46]*(1-
x[i,47])+thetahm[i,2]*(1-x[i,46])*(1-x[i,47]))-b[j,2])));
        denom[i,j] <- (1+(exp((r[i])*((thetaaf[i,1]*x[i,46]*x[i,47]+ thetahf[i,1]*(1-
x[i,46])*x[i,47]+ thetaam[i,1]*x[i,46]*(1-x[i,47])+thetahm[i,1]*(1-x[i,46])*(1-x[i,47]))-
b[j,1])+(1-r[i])*((thetaaf[i,2]*x[i,46]*x[i,47]+ thetahf[i,2]*(1-x[i,46])*x[i,47]+
thetaam[i,2]*x[i,46]*(1-x[i,47])+thetahm[i,2]*(1-x[i,46])*(1-x[i,47]))-b[j,2])))
```

```
            p[i,j] <- numer[i,j]/denom[i,j];
            x[i,j] ~ dbern(p[i,j]);
# Shadow data created and used to determine model fit
            shadow[i,j] ~ dbern(p[i,j]);
            xerr[i,j] <- pow((x[i,j]-p[i,j]),2);
            serr[i,j] <- pow((shadow[i,j]-p[i,j]),2);
            }
            xrmse[i] <- sqrt(mean(xerr[i,]));
            srmse[i] <- sqrt(mean(serr[i,]));
            count[i] <- step(srmse[i]-xrmse[i]);
}
sprop <- mean(count[]);
}
```

References

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67-91.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Bock, R.D., & Aiken, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, 46, 443-459.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2002). Item Parameter Estimation under Conditions of Test Speededness: Applications of a Mixture Rasch Model with Ordinal Constraints. *Journal of Educational Measurement*, 39, 331-348.

Bradley, J.V. (1978). Robustness? The British Journal of Mathematical & Statistical Psychology, 31, 144-152.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Casella, G., & George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46, 167-174.

Chen, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155-163.

Clauser, B.E., & Mazor, K.M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. Instructional Module for the National Council on Measurement in Education, Spring 1998.

Clauser, B., Mazor, K., & Hambleton, R.K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6(4), 269-279.

Cohen, A.S., & Bolt, D.M. (2002). A mixture model analysis of differential item functioning. Paper presented at the annual meeting of the American Educational Research Associations, New Orleans.

Cohen, J. (1988, 2nd ed.). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (ed.), *Language Proficiency and Academic Achievement*. Clevedon: Multilingual Matters.

DeAyala, R.J., Kim, Seock-Ho, Stapleton, L.M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2(3&4), 243-276.

Dorans, N.J., & Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Dorans, N.J., & Kullick, E. (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355-368.

Embretson, S.E. (1995). A measurement model for linking individual change to processes and knowledge: Application to mathematical learning. *Journal of Educational Measurement*, 32, 277-294.

Embretson, S.E. (1995). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300-326.

Embretson, S.E., & Reise, S.P. (2000). Item Response Theory for Psychologists. Mahwah, N.J.: Lawrence Erlbaum.

Fidalgo, A.M., Mellenbergh, G.J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Method of Psychological Research Online*, 5(3). Available at http://www.mpr-online.de

Fischer, G.H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.

Gierl, M.J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustration the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian Data Analysis*. New York: Chapman and Hall.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996). Introducing Markov chain Monte Carlo. In Gilks, Richardson & Spiegelhalter, *Markov chain Monte Carlo in Practice*. New York: Chapman and Hall.

Green, B.F., Crone, C.R., & Folk, V.G. (1989). A Method for Studying Differential Distractor Functioning. *Journal of Educational Measurement*, 26, 147-160.

Green K.E., & Smith, R.M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, 12, 369-381.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test item: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, 2, 313-334.

Hasting, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity*. Hillsdale, N.J.: Erlbaum.

Hu, P.G., & Dorans, N.J. (1989). *The effects of deleting differentially functioning items on equating functions and reported score distributions*. Princeton, NJ: Educational Testing Service.

Janssen, R., & DeBoeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement*, 21, 37-50.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, 27, 307-327.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18, 89-114.

Linacre, J.M., & Wright, B.D. (1989). Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions*, 3, 52-53. From the web-site: http://www.rasch.org/rmt/rmt32a.htm

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52, 443-451.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of states calculations for fast computing machines. *Journal of Chemical Physics*, 21, 1087-1091.

Meyer, J.P., Huynh, H., & Seaman, M.A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, 41, 331-344.

Mislevy, R.J., & Verhelst, N. (1990). Modeling Item Responses When Different Subjects Employ Different Solution Strategies. *Psychometrika*, 55(2), 195-215.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, 18(4), 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.

O'Neill, K.A., & McPeek, W.M. (1993). Item and Test Characteristics That are Associated with Differential Item Functioning. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Patz, R.J., & Junker, B.W. (1999). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics,* 24(2), 146-178.

Penfield, R.D. (2001). Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures. *Applied Measurement in Education,* 14(3), 235-259.

Raju, N.S. (1988). The Area Between Two Item Characteristic Curves. *Psychometrika*, 54, 495-502.

Raju, N.S., Bode, R.K., & Larsen, V.S. (1989). An empirical assessment of the Mantel-Haenszel statistic to detect differential item functioning. *Applied Measurement in Education*, 2, 1-13.

Raju, N.S. (1990). Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions. *Applied Psychological Measurement*, 14, 197-207.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Cogenhagen: Denmarks Paedagogiske Institut. (Republished in 1980 by the University of Chicago Press, Chicago).

Rivera, C., & Schmitt, A.P. (1988). *A Comparison of Hispanic and White Students' Omit Patterns on the Scholastic Aptitude Test* (Research Report No.88-44). Princeton, NJ: Educational Testing Service.

Rogers, H.J., & Swaminathan, H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, 13(2), 105-116.

Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement*, 14(3), 271-282.

Roussos, L.A., & Stout, W.F. (1996a). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement*, 33(2), 215-230.

Roussos, L.A., & Stout, W.F. (1996b). A Multidimensionality-Based DIF Analysis Paradigm. *Applied Psychological Measurement*, 20, 355-371.

Schmitt, A.P., & Bleistein, C.A. (1987). *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items* (Research Report No. 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A.P., & Dorans, N.J. (1990). Differential Item Functioning for Minority Examinees on the SAT. *Journal of Educational Measurement*, 27, 67-81.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159-194.

Sheehan, K. M., & Mislevy, R.J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement*, 27, 255-272.

Skaggs, G., & Lissitz, R.W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement*, 29(3), 227-242.

Spiegelhalter, D., Thomas, A., & Best, N. (2000). WINBUGS version 1.4 [computer program].

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WINBUGS User's manual, version 1.4*. Last accessed on May 1, 2005 from the web-site: http://www.mrc-bsu.cam.ac.uk/bugs

Standards for Educational and Psychological Testing (1999). Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Swaminathan, H., & Rogers, H.J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, 27, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Erlbaum, pg. 147-169.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, NJ: Erlbaum, pg. 67-113.

Tierney, Luke (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22, 1701-1728.

U.S. Census Bureau (1993). *The Hispanic Population in the United States: March 1993, Current Population Reports, Population Characteristics, Series P20-475*. From the web-site: http://www.census.gov/population/www/socdemo/hispanic/hispdef.html

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Weaver, C. (1994). *Reading process and practice: From sociolinguistics to whole language* (2nd ed.). Portsmouth, NH: Heinemann.

Whitely, S.E., & Schneider, L.M. (1981). Information structures on geometric analogies: A test theory approach. *Applied Psychological Measurement*, 5, 383-397.

Wollack, J.A., Cohen, A.S., & Wells, C.S. (2003). A Method for Maintaining Scale Stability in the Presence of Test Speededness. *Journal of Educational Measurement*, 40(4), 307-330.

Yamamoto, K. & Everson, H. (1996). Modeling the Effect of Test Length and Test Time on Parameter Estimation Using the HYBRID model. In J. Rost and R. Langeheine (Eds), *Applications of Latent Trait and Latent Class Models in the Social Sciences* (http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/inhalt.htm).

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, 63(1), 51-64.